A.V. Efimov, Yu.G. Zolotarev,
V.M. Terpigoreva

MATHEMATICAL ANALYSIS

Advanced Topics

Application of Some Methods of Mathematical and Functional Analysis

Mir
Publishers

Prof. Aleksandr V. Efimov, D.Sc., is the head of the Department of Higher Mathematics of the Moscow Institute of Electronic Technology. He has taught for more than thirty years (he became a lecturer in 1952).

Efimov's scientific research has been mainly in the field of the approximation properties of general orthogonal systems. He has authored many works ranging from scientific papers, monographs and books to textbooks and educational material, some of which have already been translated into English. These include: **Problem Book of Mathematics for Engineering Students. Part 1. Linear Algebra and Fundamentals of Mathematical Analysis** and **Part 2. Advanced Topics of Mathematical Analysis** (Mir Publishers, Moscow, 1984); while **Part 3. Special Courses** is in preparation.

Prof. Yuri G. Zolotarev, D.Sc., is a lecturer in the Department of Higher Mathematics of the Moscow Institute of Electronic Technology, where he has taught for thirty years. In 1954 he became a Candidate of Science and in 1974 a Doctor of the Physical and Mathematical Sciences.

He has over fifty publications. His primary scientific interests are in the fields of technical cybernetics, information theory, and applied mathematics. His other scientific writings include: "Holomorphic Functions with an Even Number of Arguments and Their Applications to Differential Equations" and "Functional Systems in Residual Classes and Their Use in Computational Systems".

A.V. Efimov, Yu.G. Zolotarev,
V.M. Terpigoreva

MATHEMATICAL ANALYSIS

Advanced Topics

Application of Some Methods of Mathematical and Functional Analysis

ИЗДАТЕЛЬСТВО
МИР

А. В. Ефимов, Ю. Г. Золотарев,
В. М. Терпигорева

Математический анализ

(Специальные разделы)

Часть 2

Применение некоторых методов математического и функционального анализа

Издательство «Высшая школа» Москва

A. V. Efimov, Yu. G. Zolotarev,
V. M. Terpigoreva

Mathematical Analysis
(Advanced Topics)

Part 2
Application of Some Methods of Mathematical and Functional Analysis

Translated from the Russian
by LEONID LEVANT

Mir Publishers Moscow

First published 1985
Revised from the 1980 Russian edition

На английском языке

© Издательство «Высшая школа», 1980
© English translation, Mir Publishers, 1985

PREFACE

This is the sequel to *Mathematical Analysis* (*Advanced Topics*) *Part 1. General Functional Series and Their Application*. It is a treatment of some of the most widespread methods of mathematical and functional analysis. The feature of this study aid is that it contains methods of functional analysis not included in textbooks for engineering students. This enables the authors to set forth approximate calculations using the methods of functional analysis and to present modern mathematical methods for solving applied problems.

The book gives an application of vector analysis to the study of various vector fields. Certain elements of functional analysis are applied to the methods of a fixed point and also to the solution of Fredholm's equations. In approximate computations, the student's attention is drawn to numerical methods of mathematical analysis which can be realized on a computer. The material contained in Chapters 3, 4, 5, 8, and 9 may be used for lectures for individual specialities.

Chapters 1, 2, 6, and 7 were written by V. M. Terpigoreva, Chapters 3, 4, and 5 by A. V. Efimov, and Chapters 8, 9, and 10 by Yu. G. Zolotarev.

Using the same layout as in Part 1, each section has its own enumeration of the theorems and formulas. When referring to a formula (theorem) in the same section, only the formula (theorem) number is given, for instance, (5). When referring to something in the same chapter, the number of the formula (theorem) is preceded by the chapter number, e.g. (2.5), and when referring to a formula (theorem) from another chapter, three numbers are given: the chapter number, the

section number, and the formula (theorem) number, for example, (3, 2, 5).

The authors are grateful to the staff of the Department of Higher Mathematics at the Moscow Institute of Electronic Technology for their useful discussions that helped to improve the manuscript.

Special thanks are due to Professors V. A. Trenogin and S. I. Pokhozhaev, and Associate Professors M. L. Krasnov, A. I. Kiselev, A. L. Pavlov, and A. M. Sedletsky for the valuable suggestions and remarks they made when reviewing the manuscript.

Invaluable assistance in editing and preparing the manuscript was rendered by A. I. Seliverstova, L. V. Lapenko, and S. A. Fomina.

The Authors

CONTENTS

Preface 5

Chapter 1. **Elements of Vector Analysis** 9
Sec. 1.1. Some Concepts of Vector Analysis 9
Sec. 1.2. Scalar Field 19
Sec. 1.3. Work Done by a Vector Field 25
Sec. 1.4. Flux of a Vector Field 36
Sec. 1.5. Divergence 45
Sec. 1.6. The Curl of a Vector Field 51

Chapter 2. **Special Kinds of Vector Fields** 64
Sec. 2.1. Potential Vector Field 64
Sec. 2.2. Solenoidal Vector Field 76
Sec. 2.3. Laplace's Vector Field 84
Sec. 2.4. Dirichlet Problem and Neumann Problem 93
Sec. 2.5. Deriving Certain Equations of Mathematical Physics 104

Chapter 3. **Certain Concepts of Functional Analysis** 111
Sec. 3.1. Statement of Problems. Hölder's and Minkowski's Inequalities 111
Sec. 3.2. Metric Spaces 121
Sec. 3.3. Completeness of Metric Spaces 132
Sec. 3.4. Contraction Mapping Principle and Its Application 138
Sec. 3.5. Compact Sets 144

Chapter 4. **Completely Continuous Operators in Normed Linear Spaces** 153
Sec. 4.1. Normed Linear Spaces 153
Sec. 4.2. Continuous and Completely Continuous Operators 156
Sec. 4.3. Schauder's Theorem and Its Application 163
Sec. 4.4. Iteration Method for Solving Fredholm's Equation 172

Chapter 5. **Self-adjoint Operators in a Hilbert Space** 178
Sec. 5.1. Basic Concepts of a Hilbert Space 178
Sec. 5.2. Self-adjoint Operators and Their Properties . . . 180
Sec. 5.3. Hilbert-Schmidt Theorem and Its Application 190

Chapter 6. **Fundamentals of the Calculus of Variations** . . . 203
Sec. 6.1. Basic Notions 203
Sec. 6.2. Extremum of a Functional 208
Sec. 6.3. Variation Problems with Fixed Boundaries . . . 215
Sec. 6.4. Variation Problems Involving a Conditional Extremum 227

Chapter 7. **Certain Methods of Solving Variation Problems** . . 241
Sec. 7.1. Variation Problems with Moving Boundaries . . . 241

Sec. 7.2. Variation Problems Involving Functions of Several Variables 250
Sec. 7.3. Connection of Variation Problems with Differential Equations 256
Sec. 7.4. Direct Methods in the Calculus of Variations . . . 263

Chapter 8. **Problems of Computation and Uniform Approximation of Functions** 283
Sec. 8.1. Errors Due to Approximate Calculations 283
Sec. 8.2. Fundamentals of the Theory of Function Approximation 288
Sec. 8.3. Polynomials of Best Approximation in Space $C[a, b]$ 293

Chapter 9. **Interpolation and Its Application to Problems of Numerical Differentiation and Integration** 304
Sec. 9.1. Interpolation 304
Sec. 9.2. Formulas for Numerical Differentiation and Integration. Error Estimates 318
Sec. 9.3. Optimization Methods. Cubature Formulas . . . 330

Chapter 10. **Numerical Methods of Solving Algebraic and Differential Equations** 339
Sec. 10.1. Systems of Linear Algebraic Equations 339
Sec. 10.2. Solving Nonlinear Equations 349
Sec. 10.3. Numerical Methods of Solving Differential Equations 351
Sec. 10.4. Net-point Method 363
Index 367

CHAPTER 1

Elements of Vector Analysis

Sec. 1.1. SOME CONCEPTS OF VECTOR ANALYSIS

1. The Vector Function of a Scalar Argument. We will say that we have a *vector function of the scalar argument* t if every value of the argument t from a certain set T is associated with a definite value of the vector $\mathbf{r}(t)$. It is possible to consider vector functions of not one but several arguments $\mathbf{r}(t_1, t_2, \ldots, t_n)$. Such concepts as the limit, continuity, partial derivatives, and some other concepts of analysis hold true for such functions.

In the Cartesian coordinate system, the specification of the vector function $\mathbf{r}(t)$ is equivalent to the specification of three scalar functions $x(t)$, $y(t)$, $z(t)$ which are its coordinates:

$$\mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j} + z(t)\,\mathbf{k}. \tag{1}$$

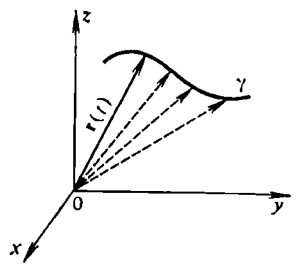


Fig. 1

Let the vector function $\mathbf{r}(t)$ be continuous on the interval $[\alpha, \beta]$, i.e. let its coordinates $x(t)$, $y(t)$, $z(t)$ be continuous. We now place the initial points of all the vectors $\mathbf{r} = \mathbf{r}(t)$, $t \in [\alpha, \beta]$, at the origin. Then the terminal points of these vectors will trace a continuous curve γ (Fig. 1). This curve is called the *hodograph*, and the vector function $\mathbf{r}(t)$ is said to be its *vector representation*. And so we write $\gamma = \{\mathbf{r}(t)\}$.

A continuously differentiable curve $\gamma = \{\mathbf{r}(t)\}$, $t \in [\alpha, \beta]$, at each point of which the derivative $\mathbf{r}'(t) \neq 0$, is said to be *smooth*.

Lemma. *At each point of the smooth curve* $\gamma = \{\mathbf{r}(t)\}$ *there exists a tangent line, and the derivative* $\mathbf{r}'(t)$ *is directed along this tangent towards increasing values of the parameter* t.

The vector $\Delta\mathbf{r}/\Delta t$ is directed along the secant $\overline{M_0M}$ of the curve $\gamma = \{\mathbf{r}(t)\}$. By the hypothesis, there exists $\lim\limits_{\Delta t \to 0} \frac{\Delta \mathbf{r}}{\Delta t} = \mathbf{r}'(t) \neq 0$, therefore $\mathbf{r}'(t)$ is directed along the tangent line to the hodograph. The vectors $\Delta\mathbf{r}$ and $\Delta\mathbf{r}/\Delta t$ lie on one and the same ray $\overline{M_0M}$. If $\Delta t > 0$, then the

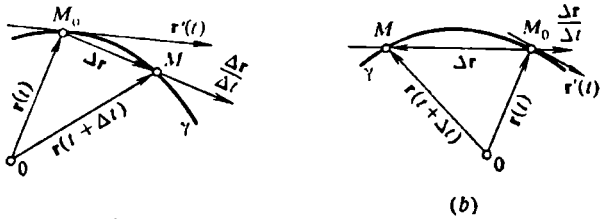


Fig. 2

vectors $\Delta\mathbf{r}$ and $\Delta\mathbf{r}/\Delta t$ are directed towards increasing values of the parameter t (Fig. 2, *a*). If $\Delta t < 0$, then the vector $\Delta\mathbf{r}$ is directed towards decreasing values of the parameter t, and the vector $\Delta\mathbf{r}/\Delta t$ is directed once again towards increasing values of the parameter t (Fig. 2, *b*).

The unit vector

$$\tau^0 = \tau^0(t) = \mathbf{r}'(t)/|\mathbf{r}'(t)| \qquad (2)$$

is also directed along the tangent towards increasing values of the parameter t. At each point of the smooth curve γ there exists a tangent on which it is possible to choose two directions: $\tau^0(t)$ and $-\tau^0(t)$.

Any unit tangent continuous on the curve γ is called the *orientation* of this curve, and a curve with a fixed orientation is said to be *oriented*. For a smooth curve there exist two orientations, one of which, defined by equality (2), is called *positive*, the other *negative*. An oriented curve γ with the initial point A and terminal point B is customarily denoted

by $\overset{\smile}{AB}$. If B is taken for the initial point and A for the terminus, then we speak of the curve $\overset{\smile}{BA}$. The curves $\overset{\smile}{AB}$ and $\overset{\smile}{BA}$ have *opposite orientations.*

A *piecewise smooth curve* is defined as a continuous curve made up of a finite number of smooth curves.

The curve $\gamma = \{\mathbf{r}(t)\}$, $t \in [\alpha, \beta]$ is said to be *closed* if its initial and terminal points coincide, that is, $\mathbf{r}(\alpha) = \mathbf{r}(\beta)$. If there are two distinct values of the parameter $t_1 \neq t_2$, $t_1, t_2 \in (\alpha, \beta)$, such that $\mathbf{r}(t_1) = \mathbf{r}(t_2)$, then the corresponding point of the curve is called the *point of self-intersection.* A closed curve having no points of self-intersection will be called a *contour.*

Let us clarify the geometrical meaning of the modulus of the differential $|\,d\mathbf{r}\,|$. We represent the derivative of the vector function (1) in the form

$$\mathbf{r}'(t) = x'(t)\,\mathbf{i} + y'(t)\,\mathbf{j} + z'(t)\,\mathbf{k},$$

and its differential in the form

$$d\mathbf{r} = \mathbf{r}'(t)\,dt = x'(t)\,dt\mathbf{i} + y'(t)\,dt\mathbf{j} + z'(t)\,dt\mathbf{k} = dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}.$$

Hence, for the modulus of the differential we obtain the expression

$$|d\mathbf{r}| = |\mathbf{r}'(t)\,dt| = \sqrt{x'^2 + y'^2 + z'^2}\,|dt|,$$

the right-hand member of which is the differential dl of the arc length of the curve. Thus, the modulus of the differential $|\,d\mathbf{r}\,|$ of the vector function $\mathbf{r}(t)$ is equal to the differential of the arc length of the hodograph, that is, $|\,d\mathbf{r}\,| = dl$.

Let us find the *direction cosines of the unit vector* $\boldsymbol{\tau}^0$ defined by equality (2) and write it in the form

$$\boldsymbol{\tau}^0 = \frac{x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k}}{\sqrt{x'^2 + y'^2 + z'^2}} = \frac{dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}}{dl}.$$

Besides, the unit vector has the following representation:

$$\boldsymbol{\tau}^0 = \mathbf{i}\cos\alpha + \mathbf{j}\cos\beta + \mathbf{k}\cos\gamma,$$

where α, β, γ are the angles made by the vector $\boldsymbol{\tau}^0$ with the x-, y- and z-axes, respectively. Comparing the two represen-

tations for τ^0, we obtain the formulas of the direction cosines of the unit vector of the tangent line:

$$\cos\alpha = \frac{x'}{\sqrt{x'^2+y'^2+z'^2}} = \frac{dx}{dl},$$
$$\cos\beta = \frac{y'}{\sqrt{x'^2+y'^2+z'^2}} = \frac{dy}{dl}, \qquad (3)$$
$$\cos\gamma = \frac{z'}{\sqrt{x'^2+y'^2+z'^2}} = \frac{dz}{dl}.$$

2. Some Information on Surfaces. The concept of the surface which is intuitively sufficiently clear can be defined with a distinct degree of generality. We will regard the surface as the image of a closed plane domain $\bar{G}$ in a continuous mapping. Such mapping can be specified in different ways. In analysis, we most frequently consider surfaces represented explicitly:

$$z = f(x, y),$$

where $f(x, y)$ is a function continuous in a closed bounded domain $\bar{G}$. A surface is regarded to be specified in a more general way if it is represented parametrically:

$$x = x(u, v), \quad y = y(u, v), \quad z = z(u, v), \qquad (4)$$

where the given functions $x(u, v)$, $y(u, v)$, $z(u, v)$ are continuous in a closed bounded plane domain $\bar{G}$. The three scalar equalities (4) can be replaced by one vector equality:

$$\mathbf{r} = \mathbf{r}(u, v) = x(u, v)\,\mathbf{i} + y(u, v)\,\mathbf{j} + z(u, v)\,\mathbf{k}. \qquad (5)$$

The surface given in vector form (5) is said to be *smooth* if the vector function $\mathbf{r} = \mathbf{r}(u, v)$ is continuously differentiable in the closed domain $\bar{G}$ and for each point of the surface the vector product $\mathbf{r}'_u \times \mathbf{r}'_v \neq 0$.

Consider the concept of orientation of a surface. Let the surface Σ have the vector representation $\mathbf{r} = \mathbf{r}(u, v)$. We fix one of the variables, for instance, $v = v_0$, then the vector function $\mathbf{r} = \mathbf{r}(u, v_0)$ will specify the curve γ_1 lying on the surface Σ. The vector function $\mathbf{r} = \mathbf{r}(u_0, v)$ determines the curve γ_2 also lying on the surface. These two curves γ_1 and γ_2 pass through the point $M = \mathbf{r}(u_0, v_0)$ belonging to the

surface Σ (Fig. 3). The vectors $\mathbf{r}'_u = \mathbf{r}'_u(u_0, v_0)$ and $\mathbf{r}'_v = \mathbf{r}'_v(u_0, v_0)$ will be respective tangents to the curves γ_1 and γ_2 at the point M. Consequently, the vectors $\mathbf{r}'_u$ and $\mathbf{r}'_v$ lie in a tangent plane to the surface Σ and the vector $\mathbf{n} = \mathbf{r}'_u \times \mathbf{r}'_v$ is directed along the normal to the surface Σ at the point M. Let $\mathbf{n}^0$ denote the unit vector of this normal:

$$\mathbf{n}^0 = \mathbf{n}^0(u, v) = \frac{\mathbf{r}'_u \times \mathbf{r}'_v}{|\mathbf{r}'_u \times \mathbf{r}'_v|}. \quad (6)$$

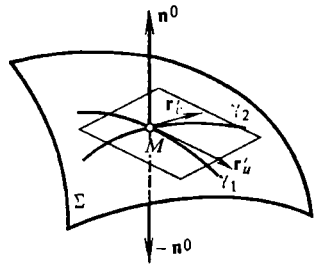


Fig. 3

At each point of the smooth surface Σ there exists a normal on which we can choose two directions: $\mathbf{n}^0$ and $-\mathbf{n}^0$ (see Fig. 3).

Any unit normal, continuous on the surface Σ is called the *orientation* of this surface, and a surface with a fixed orientation is said to be *oriented*. For a smooth surface there are two orientations one of which, determined by equality (6), is called *positive*, the other ($-\mathbf{n}^0$) *negative*.

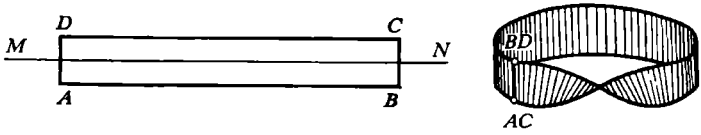


Fig. 4

Remark. If we give up the requirement of smoothness, then not all surfaces may be orientable, the Möbius strip being an example of such a surface. It is formed by taking a long rectangular strip of paper ($ABCD$) and pasting its two ends (the edge AD with the edge BC) together after giving it half a twist about the axis of symmetry MN, as it is shown in Fig. 4. In addition to being one-sided, a Möbius strip has the unusual property that it remains one piece if it is cut along the centre line (MN).

By a *piecewise smooth surface* we will understand a continuous surface made up of a finite number of smooth surfaces.

Let us find the *direction cosines of the unit normal* $\mathbf{n}^0$ determined by the equality (6) when a surface is represented in the explicit form $z = f(x, y)$. From the vector representation of this surface

$$\mathbf{r}(x, y) = x\mathbf{i} + y\mathbf{j} + f(x, y)\mathbf{k}$$

we find the derivatives

$$\mathbf{r}'_x = \mathbf{i} + f'_x\mathbf{k}, \quad \mathbf{r}'_y = \mathbf{j} + f'_y\mathbf{k},$$

write the vector product

$$\mathbf{r}'_x \times \mathbf{r}'_y = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 0 & f'_x \\ 0 & 1 & f'_y \end{vmatrix} = -f'_x\mathbf{i} - f'_y\mathbf{j} + \mathbf{k}$$

and compute its modulus: $|\mathbf{r}'_x \times \mathbf{r}'_y| = \sqrt{1 + f'^2_x + f'^2_y}$. Substituting the expressions for $\mathbf{r}'_x \times \mathbf{r}'_y$ and $|\mathbf{r}'_x \times \mathbf{r}'_y|$ into equality (6), we obtain

$$\mathbf{n}^0 = \frac{-f'_x\mathbf{i} - f'_y\mathbf{j} + \mathbf{k}}{\sqrt{1 + f'^2_x + f'^2_y}}.$$

Taking into consideration the representation of the unit vector

$$\mathbf{n}^0 = \mathbf{i}\cos\alpha + \mathbf{j}\cos\beta + \mathbf{k}\cos\gamma,$$

we finally have

$$\cos\alpha = \frac{-f'_x}{\sqrt{1 + f'^2_x + f'^2_y}}, \quad \cos\beta = \frac{-f'_y}{\sqrt{1 + f'^2_x + f'^2_y}},$$

$$\cos\gamma = \frac{1}{\sqrt{1 + f'^2_x + f'^2_y}}. \qquad (7)$$

3. Curvilinear Coordinates. Along with the Cartesian coordinate system, widely used in vector analysis are the *curvilinear coordinates*. Cylindrical and spherical coordinates, which we met in the course of analysis, are examples of curvilinear coordinates.

In each coordinate system, the position of a point in space is specified by a triple of numbers (q^1, q^2, q^3), and between the triples of numbers (q^1, q^2, q^3) and the points M of space there must be established one-to-one correspondence. An arbitrary system of coordinates (q^1, q^2, q^3) is called a *curvilinear coordinate system*. Since every point M in space can be associated with its Cartesian coordinates (x, y, z) and curvilinear coordinates (q^1, q^2, q^3), this means that between the two triples of variables x, y, z and q^1, q^2, q^3 there exists the functional relationship

$$x = x(q^1, q^2, q^3), \quad y = y(q^1, q^2, q^3), \quad z = z(q^1, q^2, q^3). \quad (8)$$

This system must be uniquely solved in the range of a triple of numbers q^1, q^2, q^3.

In terms of the curvilinear coordinates q^1, q^2, q^3, the radius vector $\mathbf{r}(M)$ of the point M will be written in the form

$$\mathbf{r}(M) = \mathbf{r}(q^1, q^2, q^3) = x(q^1, q^2, q^3)\,\mathbf{i} + y(q^1, q^2, q^3)\,\mathbf{j} + z(q^1, q^2, q^3)\,\mathbf{k}. \quad (9)$$

For a curvilinear coordinate system, let us introduce the notion of coordinate surfaces and coordinate lines. The set of points $M(q^1, q^2, q^3)$ in space for which one of the coordinates is fixed is called the *coordinate surface*. The set of points $M(q^1, q^2, q^3)$ with two coordinates fixed is called the *coordinate line*.

It is obvious that coordinate lines are intersections of coordinate surfaces. The vector equations of coordinate lines are obtained from equality (9) in which two variables are fixed. The coordinate line along which the coordinate q^ν is changed will be denoted by q^ν, $\nu = 1, 2, 3$. At each point of this line we will assume the existence of a tangent line whose vector equation has the form

$$\frac{\partial \mathbf{r}}{\partial q^\nu} = \frac{\partial x}{\partial q^\nu}\,\mathbf{i} + \frac{\partial y}{\partial q^\nu}\,\mathbf{j} + \frac{\partial z}{\partial q^\nu}\,\mathbf{k}, \quad \nu = 1, 2, 3. \quad (10)$$

For the sake of brevity, we will use the notation $\partial \mathbf{r}/\partial q^\nu = \mathbf{r}_\nu$.

The vectors $\mathbf{r}_\nu$, $\nu = 1, 2, 3$, are called the *coordinate axes*. The triple of vectors $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ form the *coordinate basis*. It is called a *local basis*.

The difference between an arbitrary curvilinear coordinate system and a Cartesian coordinate system mainly consists in that the basis vectors $\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_3$ for a curvilinear coordinate

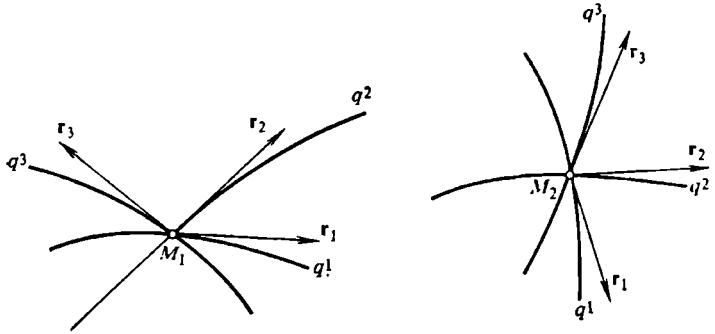


Fig. 5

system are different at distinct points in space, that is, when passing from one point (M_1) to another (M_2), the local basis $\{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3\}$ changes both in magnitude and direc-

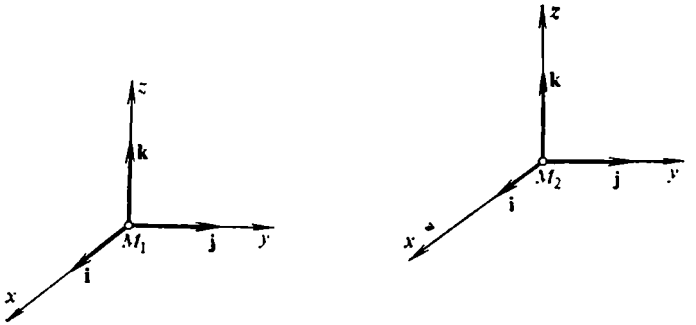


Fig. 6

tion (Fig. 5). For a Cartesian coordinate system the basis vectors $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ are the same at all points of space (Fig. 6).

A curvilinear coordinate system is said to be *orthogonal* if at any point of space the coordinate lines intersect at

right angles. Since the orthogonality of the coordinate lines means the orthogonality of the coordinate axes, the necessary and sufficient condition for a curvilinear coordinate system to be orthogonal is the equality to zero of the scalar products:

$$\mathbf{r}_\nu \cdot \mathbf{r}_s = \frac{\partial \mathbf{r}}{\partial q^\nu} \cdot \frac{\partial \mathbf{r}}{\partial q^s} = 0, \quad \nu \neq s, \quad \nu, s = 1, 2, 3,$$

or in expanded form

$$\frac{\partial x}{\partial q^\nu}\frac{\partial x}{\partial q^s} + \frac{\partial y}{\partial q^\nu}\frac{\partial y}{\partial q^s} + \frac{\partial z}{\partial q^\nu}\frac{\partial z}{\partial q^s} = 0, \quad \nu \neq s.$$

Cylindrical and spherical coordinates are examples of orthogonal curvilinear coordinate systems.

At each point M, we introduce an orthonormal basis consisting of three unit vectors:

$$\mathbf{e}_1 = \frac{\mathbf{r}_1}{|\mathbf{r}_1|}, \quad \mathbf{e}_2 = \frac{\mathbf{r}_2}{|\mathbf{r}_2|}, \quad \mathbf{e}_3 = \frac{\mathbf{r}_3}{|\mathbf{r}_3|}.$$

Although this basis $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ changes from point to point in direction, this does not prevent us from writing any vector $\mathbf{F}(M)$, given at an arbitrary point, in the form of a linear combination of vectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$:

$$\mathbf{F}(M) = F(q^1, q^2, q^3) = F^1\mathbf{e}_1 + F^2\mathbf{e}_2 + F^3\mathbf{e}_3,$$

where $F^\nu = F^\nu(q^1, q^2, q^3)$ is the projection of the vector $\mathbf{F}$ on the coordinate axis $\mathbf{r}_\nu$.

The lengths $|\mathbf{r}_\nu|$ of the basis vectors $\mathbf{r}_\nu$, $\nu = 1, 2, 3$, are usually called *Lamé's coefficients* and are denoted as $H_\nu = H_\nu(q^1, q^2, q^3)$. Thus, Lamé's coefficients are found by the formula

$$H_\nu = |\mathbf{r}_\nu| = \left|\frac{\partial \mathbf{r}}{\partial q^\nu}\right| = \sqrt{\left(\frac{\partial x}{\partial q^\nu}\right)^2 + \left(\frac{\partial y}{\partial q^\nu}\right)^2 + \left(\frac{\partial z}{\partial q^\nu}\right)^2} \quad \nu = 1, 2, 3. \quad (11)$$

Lamé's coefficients play a fundamental role when carrying out computations in curvilinear coordinate systems. The elements of length, area, and volume are expressed in terms of Lamé's coefficients; let us find them. The differential of the arc dl coincides with the length of the differential of the radius vector $|d\mathbf{r}|$, therefore the elements of the lengths of coordinate lines in the curvilinear coordinate system are

expressed in terms of Lamé's coefficients in the following way:

$$dl_\nu = |\, d\mathbf{r}_\nu \,| = H_\nu\, dq^\nu, \quad \nu = 1, 2, 3. \tag{12}$$

Consider the infinitely small rectangular parallelepiped formed by the coordinate surfaces (Fig. 7). The edges of this parallelepiped are segments of the coordinate lines having lengths $dl_\nu = H_\nu\, dq^\nu$, $\nu = 1, 2, 3$. For the areas of the faces of this parallelepiped, it is possible to write the following equalities:

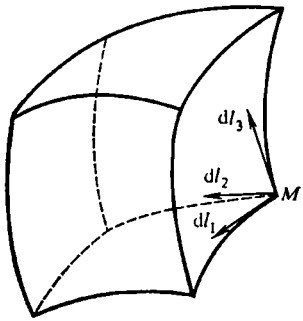


Fig. 7

$$d\sigma_{\nu s} = H_\nu H_s\, dq^\nu\, dq^s, \quad \nu \neq s, \quad \nu, s = 1, 2, 3. \tag{13}$$

In the same way, we obtain the expression (also in terms of Lamé's coefficients) for the volume of the infinitely small parallelepiped:

$$dv = dl_1\, dl_2\, dl_3 = H_1 H_2 H_3\, dq^1\, dq^2\, dq^3.$$

Formulas (12), (13), and (14) enable us to treat the curvilinear coordinate system q^ν, $\nu = 1, 2, 3$, in the given space R as a Cartesian coordinate system q^ν, $\nu = 1, 2, 3$, but in another space $\tilde{R}$. Here, the infinitely small curvilinear parallelepiped with sides dl_1, dl_2, dl_3 can be regarded as the image of the rectangular parallelepiped with sides dq^1, dq^2, dq^3.

Later, we shall need Lamé's coefficients for cylindrical and spherical coordinate systems. The cylindrical coordinates ρ, φ, and z ($\rho \geqslant 0$, $0 \leqslant \varphi < 2\pi$, $-\infty < z < \infty$) are related with the Cartesian coordinates x, y, and z by the formulas

$$x = \rho \cos \varphi, \quad y = \rho \sin \varphi, \quad z = z. \tag{15}$$

Lamé's coefficients for the cylindrical coordinate system have the following values:

$$H_\rho = \sqrt{\cos^2 \varphi + \sin^2 \varphi} = 1,$$
$$H_\varphi = \sqrt{\rho^2 \sin^2 \varphi + \rho^2 \cos^2 \varphi} = \rho, \quad H_z = 1. \tag{16}$$

The spherical coordinates r, θ, and φ ($r \geqslant 0$, $0 \leqslant \theta \leqslant \pi$, $0 \leqslant \varphi < 2\pi$) and the Cartesian coordinates x, y, and z are related in the following way:

$$x = r \cos \varphi \sin \theta, \quad y = r \sin \theta \sin \varphi, \quad z = r \cos \theta. \tag{17}$$

Lamé's coefficients for the spherical coordinate system have the following values:

$$H_r = \sqrt{\cos^2 \varphi \sin^2 \theta + \sin^2 \varphi \sin^2 \theta + \cos^2 \theta} = 1,$$
$$H_\theta = \sqrt{r^2 \cos^2 \varphi \cos^2 \theta + r^2 \sin^2 \varphi \cos^2 \theta + r^2 \sin^2 \theta} = r, \tag{18}$$
$$H_\varphi = \sqrt{r^2 \sin^2 \varphi \sin^2 \theta + r^2 \cos^2 \varphi \sin^2 \theta} = r \sin \theta.$$

Sec. 1.2. SCALAR FIELD

1. The Concept of Scalar and Vector Fields. The notion of the field underlies many ideas of contemporary physics. From the physical point of view, field is defined as a certain part of space in which some physical phenomenon, we are interested in, takes place. Putting aside the physical meaning of the field, we shall study the so-called mathematical theory of field.

We shall say that a *field* is specified in a certain domain Ω if to every point of this domain there corresponds a definite value of some quantity, numerical or vector. If at each point of the domain Ω the given quantity takes on numerical values, then the field is called *scalar*, and if at each point of the domain Ω a vector is given, then such a field is termed the *vector field*. The specification of a scalar field means that at every point $M \in \Omega$ having the radius vector $\mathbf{r} = \mathbf{r}(M)$ a scalar function $f(M) = f(\mathbf{r})$ is defined, whereas the specification of a vector field is characterized by defining a vector function $\mathbf{F}(M) = \mathbf{F}(\mathbf{r})$.

A temperature field, illumination intensity field, electric charge density field, mass density field are a few examples of scalar fields. Thus, if each point M of a heated body is associated with its temperature $T(M)$, then it forms a field of temperatures inside the heated body. A source of light creates a scalar field of illumination intensity and each point M is associated with illumination at this point. Each point M of a domain with continuously distributed electric

charges can be associated with the density of electric charges $\rho(M)$, and, thus, we obtain a scalar field of density of electric charges. A continuously distributed mass in a certain region forms a scalar field of mass density: each point M is associated with the mass density $\rho(M)$ at this point.

Here are several examples of vector fields: velocity field of a moving fluid, gravitational field, electrostatic field. If in some domain filled with fluid running with a certain velocity which, generally speaking, differs from point to point, then every point M can be associated with the velocity vector $\mathbf{V} = \mathbf{V}(M)$, and we obtain a vector field of velocities of a flowing fluid. If some mass is distributed in a certain domain, then a material point with unit mass placed at a given point M is acted upon by the gravitational force $\mathbf{F}(M)$ which forms a vector field of gravitational forces or a gravitational field. Electric charges distributed in a certain domain act with a definite force $\mathbf{F}(M)$ upon the unit charge placed at the point M. These forces generate a vector field called the electrostatic field.

A scalar or a vector field is called *stationary* if the quantity under consideration depends only on the position of a point in space, but is independent of time. If the quantity in question also depends on time, then the field is called *nonstationary*.

For instance, the above mentioned fields of temperature distribution and fluid velocities can be both stationary and nonstationary.

2. Level Lines and Level Surfaces. For the sake of obviousness, use is made of the graphical representation of a scalar field. A *level surface of a scalar field* $f(M)$ is defined as the locus of points M at which the field $f(M)$ has the given constant value c. A nonstationary field should be considered at a specific instant of time. Consider some coordinate system in which the equation of the level surface has the form

$$f(M) = c.$$

Level surfaces corresponding to different c's fill the entire domain in which the field is defined, and no two surfaces $f(M) = c_1$ and $f(M) = c_2$, $c_1 \neq c_2$, have common points. Representation of all level surfaces with appropriate values

of c is equivalent to specifying the field $f(M)$ itself. Mutual positions of level surfaces give an obvious idea of a certain scalar field. The places at which the surfaces come closer point to a rapid change of the function $f(M)$, a slow change of the function $f(M)$ being indicated by the places where surfaces are spaced out.

Example 1°. The potential of an electrostatic field forms a scalar field. Find its level surfaces.

This scalar field is convenient to be considered in the spherical coordinate system. It is characterized by the scalar function e/r, where e is the quantity of electric charge placed at the origin of coordinates, and r is the distance from the origin to the point under consideration. Here, the level surfaces are the spheres $r = c$. On each sphere, the potential will be inversely proportional to the radius of the sphere.

If a scalar field is defined in a plane domain, then, instead of level surfaces, *level lines* are considered. With the aid of level lines, we usually represent temperature distribution (isotherms), pressure distribution (isobars), relief of a terrain on a topographical map (contour lines).

3. Directional Derivative. Level surfaces enable us to judge of the rate of change of the scalar field $f(M)$ in a given direction only qualitatively. The quantitative characteristic of the rate of change of the field $f(M)$ is given by the directional derivative. Let us recall this notion.

Let a scalar field $f(M)$ be defined in a domain Ω. Consider the point $M \in \Omega$ and some fixed direction specified by the unit vector $\mathbf{s}^0$. Through the point M, we draw a straight line s parallel to the vector $\mathbf{s}^0$, take a point M_1 on it, and write the ratio

$$\frac{f(M_1) - f(M)}{MM_1}.$$

If there exists the limit of this ratio as the point M_1 moving along the straight line s approaches M, then it is called the *(directional) derivative of the scalar field* $f(M)$ in the direction s at the point M and is denoted df/ds. Thus, we obtain the definition of the directional derivative:

$$\frac{df}{ds} = \lim_{M_1 \to M} \frac{f(M_1) - f(M)}{MM_1}. \tag{1}$$

$M_1 \to M$ along the straight line s.

Such a definition of the derivative does not require the choice of a coordinate system. The derivative df/ds characterizes the rate of change of the field $f(M)$ in the direction s. In the Cartesian coordinate system, df/ds is computed by the formula

$$\frac{df}{ds} = \frac{\partial f}{\partial x}\cos\alpha + \frac{\partial f}{\partial y}\cos\beta + \frac{\partial f}{\partial z}\cos\gamma, \qquad (2)$$

where α, β, γ are the angles made by the vector s with the coordinate axes.

4. Gradient. The gradient is an important characteristic of a scalar field which makes it possible to describe this field analytically. The notion of gradient was introduced in the course of analysis. In the Cartesian coordinate system, the *gradient of the scalar field* $f(M)$ is determined by the formula

$$\operatorname{grad} f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}. \qquad (3)$$

This determination holds true only for the Cartesian coordinate system. Let us recall the basic properties of the gradient.

(a) *The derivative in any direction* s *is equal to the scalar product of the gradient by the unit vector of this direction*, that is,

$$\frac{df}{ds} = \operatorname{grad} f \cdot \mathbf{s}^0 = |\operatorname{grad} f| \cos(\widehat{\mathbf{s}, \operatorname{grad} f}) \qquad (4)$$

or

$$\frac{df}{ds} = \operatorname{pr}_{\mathbf{s}} \operatorname{grad} f(M).$$

Using this property, we can give the definition of gradient independent of the coordinate system.

The *gradient of the scalar field* $f(M)$ *at the point* M is the vector whose projection on the direction s is equal to the derivative of the field f in this direction, that is,

$$\operatorname{pr}_{\mathbf{s}} \operatorname{grad} f(M) = (\operatorname{grad} f)_{\mathbf{s}} = \frac{df}{ds}.$$

(b) *The directional derivative* df/ds *attains its greatest value in the direction of* grad f, *and the greatest value is equal to*

the modulus of the gradient, that is,

$$\max \frac{df}{ds} = |\operatorname{grad} f| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \left(\frac{\partial f}{\partial z}\right)^2}.$$

(c) *The vector* grad f *is normal to the level surface and directed towards increasing values of the function* $f(M)$.

Hence it follows immediately that if $f(M) = c$ is a level surface in a scalar field, and $\mathbf{n}^0(M)$ is the unit normal at the point M, then

$$\mathbf{n}^0 = \pm \frac{\operatorname{grad} f}{|\operatorname{grad} f|}, \tag{5}$$

where the sign is chosen depending on the orientation of the surface.

(d) *If* grad $f = 0$ *in the domain* Ω, *then* $f \equiv \text{const}$ *in* Ω.

5. Gradient in Orthogonal Curvilinear Coordinate System. Using the definition of the gradient, independent of the coordinate system, it is possible to show that the *gradient in an orthogonal curvilinear coordinate system* is computed by the formula

$$\operatorname{grad} f = \frac{1}{H_1}\frac{\partial f}{\partial q^1}\mathbf{e}_1 + \frac{1}{H_2}\frac{\partial f}{\partial q^2}\mathbf{e}_2 + \frac{1}{H_3}\frac{\partial f}{\partial q^3}\mathbf{e}_3. \tag{6}$$

Lamé's coefficients for the cylindrical coordinate system are found by formula (16) from Sec. 1.1. Substituting them into (6), we obtain the expression for the gradient in the cylindrical coordinate system:

$$\operatorname{grad} f = \frac{\partial f}{\partial \rho}\mathbf{e}_\rho + \frac{1}{\rho}\frac{\partial f}{\partial \varphi}\mathbf{e}_\varphi + \frac{\partial f}{\partial z}\mathbf{e}_z. \tag{7}$$

Using formulas (18) given at the close of the preceding section, we find the expression for the gradient in the spherical coordinate system:

$$\operatorname{grad} f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{e}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \varphi}\mathbf{e}_\varphi. \tag{8}$$

Let us dwell on a particular case of formula (8). Suppose the scalar field $f(M) = f(r)$ is determined by a function depending only on the modulus $r = |\mathbf{r}|$ of the radius vector $\mathbf{r}$. In this case we obtain the following convenient formula

for computing the gradient:

$$\operatorname{grad} f(r) = \frac{df}{dr}\,\mathbf{e}_r = f'(r)\,\frac{\mathbf{r}}{r}. \tag{9}$$

Example 2°. Find grad r^3, where $\mathbf{r} = \sqrt{x^2 + y^2 + z^2}$. By formula (9), we have

$$\operatorname{grad} r^3 = 3r^2\,\frac{\mathbf{r}}{r} = 3r\mathbf{r} = 3\sqrt{x^2+y^2+z^2}\,(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}).$$

Example 3°. In the cylindrical coordinate system, a scalar field is given by the function $f = 1/\rho$, where $\rho = \sqrt{x^2 + y^2}$. Find grad f.

Making use of formula (7), we get

$$\operatorname{grad}\frac{1}{\rho} = -\frac{1}{\rho^2}\,\mathbf{e}_\rho = \frac{-\rho}{\rho^3} = -\frac{x\mathbf{i}+y\mathbf{j}}{\sqrt{(x^2+y^2)^3}}.$$

Gradient is widely used in various applied problems of physics and engineering. Let us consider one such problem.

Example 4°. Find the law of refraction of light at the boundary surface γ of two homogeneous media.

Let λ be the index of refraction of the second medium with respect to the first, α the angle of incidence of the ray, and β the angle of refraction. The index of refraction of the second medium with respect to the first shows that light propagates in the first medium with a velocity λ times greater than in the second. It is known from physics that the ray MPN (Fig. 8) must have such a shape that the time during which light covers the distance $MP + PN$ is minimum. Let us introduce the following notation: $\overline{MP} = r_1$ and $\overline{NP} = r_2$. The time during which light covers the distance $MP + PN$ is proportional to the expression $f(P) = r_1 + \lambda r_2$, and therefore the problem is reduced to finding the minimum of the function $f(P)$. At the point of minimum there must be fulfilled the equality $df/ds = 0$, where s is the direction of the tangent to γ at the point P. The derivative df/ds is found

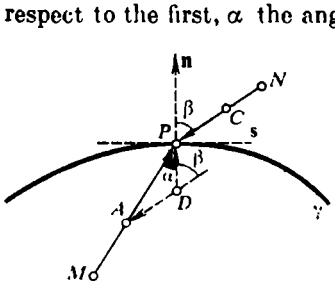


Fig. 8

by formula (4), and, thus, at the point of minimum we have

$$\frac{df}{ds} = |\operatorname{grad} f| \cos(\widehat{s \operatorname{grad} f}) = 0.$$

This equality means that grad f is directed along the normal $\mathbf{n}$ to the curve γ. Further, since $f = r_1 + \lambda r_2$, we get

$$\operatorname{grad} f = \operatorname{grad} r_1 + \lambda \operatorname{grad} r_2 = \frac{\mathbf{r}_1}{r_1} + \lambda \frac{\mathbf{r}_2}{r_2}.$$

Thus, the vector $\frac{\mathbf{r}_1}{r_1} + \lambda \frac{\mathbf{r}_2}{r_2}$ is in the direction of the normal $\mathbf{n}$ to the curve γ: $\mathbf{r}_1/r_1 = \overrightarrow{AP}$, $\lambda(\mathbf{r}_2/r_2) = \overrightarrow{CP} = \overrightarrow{DA}$ (see Fig. 8). Hence we find the lengths $AP = |\overrightarrow{AP}| = 1$ and $DA = |\overrightarrow{DA}| = \lambda$. By the sine theorem, from the triangle PAD we have

$$\frac{AP}{\sin(\pi - \beta)} = \frac{DA}{\sin \alpha}, \quad \text{or} \quad \frac{1}{\sin \beta} = \frac{\lambda}{\sin \alpha}.$$

Hence we obtain the equality $\sin \alpha = \lambda \sin \beta$, expressing the law of refraction of light on the boundary of two media.

Sec. 1.3. WORK DONE BY A VECTOR FIELD

1. Vector Lines. Geometrically, a vector field is characterized by vector lines.

A *vector line* (or a *line of force*) *of a vector field* $\mathbf{F}(M)$ is a curve at whose every point the direction of its tangent τ coincides with that of the vector of the field (Fig. 9).

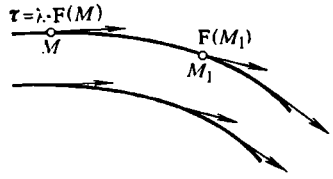


Fig. 9

For instance, in the above considered stationary velocity field of a fluid flow (see Sec. 1.2) the vector lines are the so-called stream lines (or flow lines) which serve as the trajectories of motion of the particles of the fluid.

As any curve, a vector line can be characterized by its equation $\mathbf{r}(t)$ which depends on the coordinate system chosen. Let us derive the differential equation of a vector line in the Cartesian coordinate system. For the sake of convenience,

let us denote the projections of the vector F on the coordinate axes Ox, Oy, and Oz as follows:

$$F_x = F^1(x, y, z), \quad F_y = F^2(x, y, z), \quad F_z = F^3(x, y, z),$$

respectively. Then the vector field is specified by the equality $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$. The vector $d\mathbf{r} = \mathbf{r}'(t)\,dt = dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}$ is directed as a tangent line to $\mathbf{r}(t)$ and, by the definition of a vector line, is collinear with the vector of the field F. From the condition of collinearity of vectors, we obtain the differential equation of the vector lines in the Cartesian coordinate system:

$$\frac{dx}{F_x} = \frac{dy}{F_y} = \frac{dz}{F_z}. \tag{1}$$

To determine the vector lines, we have to solve the system of differential equations (1). We will assume that the projections F_x, F_y, and F_z are continuous and possess continuous partial derivatives. From the theorem of existence and uniqueness of the solution of the system of differential equations it is known that if the vector $\mathbf{F}(M)$ at the point M is different from zero, then through this point there passes a unique vector line which is the solution of equation (1). If $\mathbf{F}(M) = 0$, then all the denominators in equation (1) vanish, and there may pass through this point either infinitely many vector lines or none at all.

If the field $\mathbf{F} = F^1\mathbf{e}_1 + F^2\mathbf{e}_2 + F^3\mathbf{e}_3$ is considered in the curvilinear coordinate system q^ν, $\nu = 1, 2, 3$, then the vector lines are found from the condition of collinearity of the vectors F and $d\mathbf{r} = H_1\,dq^1\,\mathbf{e}_1 + H_2\,dq^2\,\mathbf{e}_2 + H_3\,dq^3\,\mathbf{e}_3$; and this leads to the system of differential equations

$$\frac{H_1\,dq^1}{F^1} = \frac{H_2\,dq^2}{F^2} = \frac{H_3\,dq^3}{F^3}. \tag{2}$$

In particular, in cylindrical coordinates equation (2) is written in the form

$$\frac{d\rho}{F_\rho} = \frac{\rho\,d\varphi}{F_\varphi} = \frac{dz}{F_z}, \tag{3}$$

and in spherical coordinates in the form

$$\frac{dr}{F_r} = \frac{r\,d\theta}{F_\theta} = \frac{r\sin\theta\,d\varphi}{F_\varphi}. \tag{4}$$

Example 1°. Determine the vector lines of magnetic intensity of the field set up by an electric current of strength I flowing in an infinitely long straight wire.

First of all, we find the vector $\mathbf{H}$ of magnetic field strength. Let the z-axis be directed along the wire, and let the current flow in the positive direction of the z-axis. According to the Biot-Savart law, the wire element dt produces at the point $M(x, y, z)$ the intensity

$$d\mathbf{H} = \frac{I}{r_1^3}(dt\mathbf{k} \times \mathbf{r}_1),$$

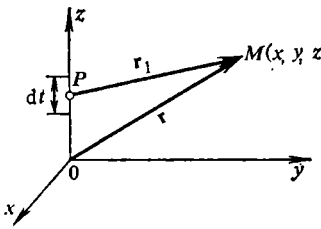


Fig. 10

where $\mathbf{r}_1 = \overrightarrow{PM}$, $r_1 = |\overrightarrow{PM}|$ is the length of the vector $\mathbf{r}_1$, $P(0, 0, t)$ is a point belonging to the wire element dt (Fig. 10). Integrating along the z-axis, we find the magnetic field strength at the point M:

$$\mathbf{H} = \mathbf{H}(M) = I \int_{-\infty}^{\infty} \frac{\mathbf{k} \times \mathbf{r}_1}{r_1^3}\, dt. \qquad (5)$$

Now $r_1 = \sqrt{x^2 + y^2 + (z - t)^2}$, since $\mathbf{r}_1 = \overrightarrow{PM} = x\mathbf{i} + y\mathbf{j} + (z - t)\mathbf{k}$. Computing, in addition, the vector product

$$\mathbf{k} \times \mathbf{r}_1 = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & 1 \\ x & y & z - t \end{vmatrix} = y\mathbf{i} + x\mathbf{j}$$

and substituting these expressions under the integral sign into equality (5), we obtain

$$\mathbf{H} = I \int_{-\infty}^{\infty} \frac{-y\mathbf{i} + x\mathbf{j}}{\sqrt{(x^2 + y^2 + (z - t)^2)^3}}\, dt.$$

The last integral is evaluated with the aid of the substitution $t - z = \sqrt{x^2 + y^2} \tan \alpha$, which leads it to the form

$$\mathbf{H} = \frac{I(-y\mathbf{i} + x\mathbf{j})}{x^2 + y^2} \int_{-\pi/2}^{\pi/2} \cos \alpha\, d\alpha = \frac{2I}{x^2 + y^2}(-y\mathbf{i} + x\mathbf{j}).$$

To obtain the equation of the lines of force, we have to solve the system of differential equations (1), which in this case takes the form

$$\frac{dx}{-2Iy/(x^2+y^2)} = \frac{dy}{2Ix/(x^2+y^2)} = \frac{dz}{0},$$

or, on reducing by $2I/(x^2+y^2)$,

$$\frac{dx}{-y} = \frac{dy}{x} = \frac{dz}{0}.$$

Hence we obtain the system of equations

$$x\,dx + y\,dy = 0, \quad dz = 0.$$

Solving this system, we find a family of equations of the vector lines of the magnetic field intensity:

$$x^2 + y^2 = R^2, \quad z = c.$$

The constants c and R are determined from the condition of passing of a vector line through a definite point $M_0(x_0, y_0, z_0)$. Through any point not lying on the z-axis there passes a unique vector line which is a circle lying in the plane parallel to the xy-plane.

2. Line Integrals of the Second Kind. Let us introduce the notion of the line integral of the second kind. We shall do it with the aid of the line integral of the first kind $\int_\gamma f(M)\,dl$, where γ is a smooth curve at each point of which a continuous scalar function $f(M)$ is defined. Let γ be a smooth oriented curve, $\tau^0 = \tau^0(M)$ the unit vector of the tangent line to this curve at the point M, and $\mathbf{F}(M)$ a vector function defined and continuous on γ.

The line integral of the first kind, when $f = \mathbf{F}\cdot\tau^0$ is a scalar product of the vector $\mathbf{F}$ by the unit vector of the tangent τ^0, that is, the integral

$$\int_\gamma \mathbf{F}\cdot\tau^0\,dl \tag{6}$$

will be called the *line integral of the second kind* of the vector $\mathbf{F}$ over the curve γ.

Using the relationship $\tau^0\,dl = d\mathbf{r}$, we write the line integral of the second kind in another form

$$\int_{\gamma} \mathbf{F}\cdot d\mathbf{r} = \int_{\gamma} \mathbf{F}\cdot\tau^0\,dl. \tag{7}$$

With the aid of the relationships

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k},$$
$$d\mathbf{r} = dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k},$$
$$\tau^0 = \cos\alpha\mathbf{i} + \cos\beta\mathbf{j} + \cos\gamma\mathbf{k},$$

we write equality (7) in the Cartesian coordinate system:

$$\int_{\gamma} F_x\,dx + F_y\,dy + F_z\,dz = \int_{\gamma} (F_x\cos\alpha + F_y\cos\beta + F_z\cos\gamma)\,dl. \tag{8}$$

When the orientation of the curve is changed, the line integral of the second kind changes its sign, that is,

$$\int_{\overset{\smile}{AB}} \mathbf{F}\cdot d\mathbf{r} = -\int_{\overset{\smile}{BA}} \mathbf{F}\cdot d\mathbf{r}. \tag{9}$$

Indeed, if τ^0 is the unit vector of the tangent line to the curve $\overset{\smile}{AB}$, then, when the orientation of the curve is changed, its unit vector will be $\tau_1^0 = -\tau^0$. Using equality (7), we get

$$\int_{\overset{\smile}{BA}} \mathbf{F}\cdot d\mathbf{r} = \int_{\overset{\smile}{BA}} \mathbf{F}\cdot\tau_1^0\,dl = -\int_{\overset{\smile}{AB}} \mathbf{F}\cdot\tau^0\,dl = -\int_{\overset{\smile}{AB}} \mathbf{F}\cdot d\mathbf{r}.$$

Line integrals of the second kind can also be determined for piecewise smooth curves. If Γ is a piecewise smooth curve, that is, it is representable as a union of a finite number of smooth curves γ_k, $k = 1, 2, \ldots, n$, $\Gamma = \bigcup_{k=1}^{n} \gamma_k$, then, by definition, we get

$$\int_{\Gamma} \mathbf{F}\cdot d\mathbf{r} = \sum_{k=1}^{n} \int_{\gamma_k} \mathbf{F}\cdot d\mathbf{r}.$$

3. Work Done by a Vector Field. Let us give the physical interpretation of the line integral (6). If a continuous field of forces $\mathbf{F}(M)$ is specified in a certain domain Ω, then it is known that, as a material point moves along a smooth oriented curve γ, the field will do some work A. In order to determine this work, let us partition the curve γ into n arcs by the division points $B = M_0, M_1, \ldots, M_k, M_{k+1}, \ldots, M_n = C$ (Fig. 11). Let P_k be an arbitrary point of the arc $M_k M_{k+1}$, and let $\tau^0(P_k)$ be the unit vector of the tangent line to the curve γ at this point. If Δl_k is the arc length

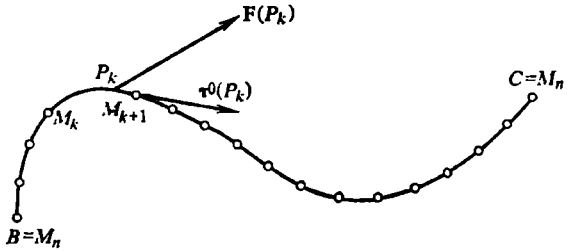


Fig. 11

of the curve $M_k M_{k+1}$, then the work ΔA_k done on the section $M_k M_{k+1}$ can be computed approximately with the aid of the scalar product:

$$\Delta A_k \approx \mathbf{F}(P_k) \cdot \tau^0(P_k)\, \Delta l_k.$$

For the work A done over the entire curve $\overset{\frown}{BC}$ it is natural to take the limit

$$A = \lim_{\max \Delta l_k \to 0} \sum_{k=1}^{n} \mathbf{F}(P_k) \cdot \tau^0(P_k)\, \Delta l_k.$$

This limit, provided it exists, is the line integral of the first kind of the scalar product $\mathbf{F}(M) \cdot \tau^0(M)$, that is, the line integral of the second kind (6). Thus, the work A done to displace a material point in a continuous field of forces is expressed by the line integral of the second kind:

$$A = \int_{\gamma} \mathbf{F} \cdot \tau^0 \, dl = \int_{\gamma} \mathbf{F} \cdot d\mathbf{r}. \tag{10}$$

If the curve γ has the vector representation $\mathbf{r} = \mathbf{r}(t)$, $t \in [\alpha, \beta]$, then the work A is computed by the formula

$$\int_{\gamma} \mathbf{F} \cdot d\mathbf{r} = \int_{\alpha}^{\beta} \mathbf{F}(M(t)) \cdot \mathbf{r}'(t)\, dt. \tag{11}$$

Substituting the value of $\mathbf{r}'(t)$ into this formula, we obtain the formula for computing work in the Cartesian coordinate system:

$$\int_{\gamma} \mathbf{F} \cdot d\mathbf{r} = \int_{\alpha}^{\beta} [F_x(x(t), y(t), z(t)) x'(t) + F_y(x(t), y(t), z(t)) y'(t) + F_z(x(t), y(t), z(t)) z'(t)]\, dt. \tag{12}$$

If the field $\mathbf{F}$ is plane and the curve γ is represented explicitly: $y = y(x)$, $a \leqslant x \leqslant b$, then the work (10) is computed by the formula

$$\int_{\gamma} \mathbf{F} \cdot d\mathbf{r} = \int_{a}^{b} [F_x(x, y(x)) + F_y(x, y(x)) y'(x)]\, dx.$$

Example 1°. Compute the work of the field of forces $\mathbf{F} = x^2\mathbf{i} + y\mathbf{j} + yz\mathbf{k}$ along the first turn of the helical line $x = \cos t$, $y = \sin t$, $z = 2t$, $0 \leqslant t \leqslant 2\pi$.

Applying formula (12), we find

$$A = \int_{\gamma} \mathbf{F} \cdot d\mathbf{r} = \int_{0}^{2\pi} (-\cos^2 t \sin t + \sin t \cos t - 4t \sin t)\, dt$$
$$= \frac{1}{3}\cos^3 t \Big|_0^{2\pi} - \frac{1}{4}\cos^2 t \Big|_0^{2\pi} + 4t \cos t \Big|_0^{2\pi} - 4 \sin t \Big|_0^{2\pi} = 8\pi.$$

Example 2°. Compute the work of the field $\mathbf{F} = \mathbf{r}/r^3$ along the radius vector $\mathbf{r}$ from the point $M_1(r_1)$ to the point $M_2(r_2)$.

The equality $\mathbf{r}^2 = r^2$ implies the equality $2\mathbf{r}\, d\mathbf{r} = 2r\, dr$, therefore we have

$$A = \int_{\gamma} \frac{\mathbf{r}\, d\mathbf{r}}{r^3} = \int_{r_1}^{r_2} \frac{dr}{r^2} = -\frac{1}{r}\Big|_{r_1}^{r_2} = \frac{1}{r_1} - \frac{1}{r_2}.$$

Example 3°. Show that the work of the field **F** along any vector line of this field is different from zero.

Let γ be a vector of the field **F**. Then at each point M of the curve γ the tangent to γ is in the direction of the vector $\mathbf{F}(M)$. Hence we obtain the scalar product $\mathbf{F}\cdot\tau^0 = |\mathbf{F}| > 0$, and therefore

$$\int_{\gamma} \mathbf{F}\cdot\tau^0\, dl = \int_{(\gamma)} |\mathbf{F}|\, dl > 0,$$

here the curve γ may be closed.

4. Green's Formula. This formula, which is widely applied in vector analysis, establishes a relation between the work of a plane vector field **F** round the contour γ and the double integral taken over the domain G bounded by this contour.

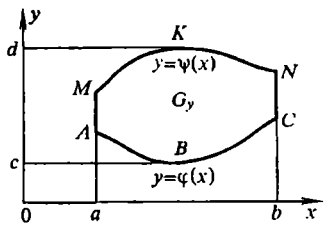


Fig. 12

Let us first introduce the notions of the positive and negative orientations of the contour γ. The orientation of the contour γ is said to be *positive* if when tracing the contour γ corresponding to increasing values of the parameter the domain G is kept on the left (such tracing is usually called the anticlockwise tracing of the contour), otherwise it is *negative*. A positively oriented contour will be denoted by γ^+, and a negatively oriented contour by γ^-. These notions are defined not strictly, but they will help us to get a geometrical obviousness of the things to be considered on the next pages.

We are going to prove Green's formula for simple domains G. Here, the plane domain G_y is regarded as *simple* with respect to the y-axis if its boundary consists of the graphs of two continuous on $[a, b]$ functions $y = \varphi(x)$ and $y = \psi(x)$, $\varphi(x) \leqslant \psi(x)$, and possibly of two segments of the straight lines $x = a$ and $x = b$ parallel to the y-axis (Fig. 12).

A domain, which is simple with respect to the x-axis, is defined in a similar way. The domain G, bounded by a piecewise smooth contour γ is called *simple* if it can be sepa-

rated into a finite number of subdomains, simple with respect to both coordinate axes.

Theorem 1. *If the domain G is simple and the vector function* $\mathbf{F} = F_x(x, y)\,\mathbf{i} + F_y(x, y)\,\mathbf{j}$, *and the partial derivatives* $\partial F_x/\partial y$, $\partial F_y/\partial x$ *are continuous in the closed domain* $\overline{G} = G + \gamma$, *the Green's formula holds for* G:

$$\iint_G \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\,dy = \oint_{\gamma+} F_x\,dx + F_y\,dy. \qquad (13)$$

Let us first prove the theorem for the domain G which is simple simultaneously with respect to both coordinate axes (Ox and Oy in Fig. 12). Regarding G as a simple domain with respect to the y-axis, that is, $G = G_y$, we reduce the double integral $\iint_G \frac{\partial F_x}{\partial y}\,dx\,dy$ to an iterated integral:

$$\iint_{G_y} \frac{\partial F_x}{\partial y}\,dx\,dy = \int_a^b dx \int_{\varphi(x)}^{\psi(x)} \frac{\partial F_x}{\partial y}\,dy$$

$$= \int_a^b F_x(x, \psi(x))\,dx - \int_a^b F_x(x, \varphi(x))\,dx. \qquad (14)$$

The definite integral $\int_a^b F_x(x, \psi(x))\,dx$ can be replaced by the line integral over the curve $\overset{\frown}{MKN}$, i.e.

$$\int_a^b F_x(x, \psi(x))\,dx = \int_{\overset{\frown}{MKN}} F_x(x, y)\,dx = -\int_{\overset{\frown}{NKM}} F_x(x, y)\,dx.$$

Analogously, the integral $\int_a^b F_x(x, \varphi(x))\,dx$ can be replaced by the line integral over the curve $\overset{\frown}{ABC}$, i.e.

$$\int_a^b F_x(x, \varphi(x))\,dx = \int_{\overset{\frown}{ABC}} F_x(x, y)\,dx.$$

On the line segments MA and CN we have $dx = 0$, therefore the integrals $\int\limits_{\overset{\frown}{MA}} F_x\,dx$ and $\int\limits_{\overset{\frown}{CN}} F_x\,dx$ are also equal to zero. Then equality (14) can be rewritten in the following way:

$$\iint\limits_{G_y} \frac{\partial F_x}{\partial y}\,dx\,dy = -\int\limits_{\overset{\frown}{AC}} F_x\,dx - \int\limits_{\overset{\frown}{CN}} F_x\,dx - \int\limits_{\overset{\frown}{NM}} F_x\,dx - \int\limits_{\overset{\frown}{MA}} F_x\,dx = -\oint\limits_{\gamma} F_x\,dx.$$

Regarding the domain G as a simple domain with respect to the x-axis, that is, assuming $G = G_x$ and reasoning in the same way as in the case $G = G_y$, we obtain the formula

$$\iint\limits_{G_x} \frac{\partial F_y}{\partial x}\,dx\,dy = \oint\limits_{\gamma} F_y\,dy.$$

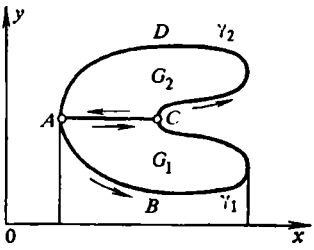


Fig. 13

Adding together the last two equalities, we find Green's formula (13) for the case under consideration.

Now, it is easy to solve the theorem for the general case.

Let $G = \bigcup\limits_{i=1}^{n} G_i$, where G_i, $i = 1, 2, \ldots, n$, are domains, simple with respect to the coordinate axes. For the sake of simplicity, let us consider the case when $G = G_1 \cup G_2$ (Fig. 13). The domain G_1 is bounded by the contour $\gamma_1 = \overset{\frown}{ABCA}$, and the domain G_2 by the contour $\gamma_2 = \overset{\frown}{ACDA}$. For either of the domains G_1 and G_2, Green's formula holds true, therefore we may write:

$$\iint\limits_{G_1} \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\,dy + \iint\limits_{G_2} \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\,dy = \oint\limits_{\gamma_1} F_x\,dx + F_y\,dy + \oint\limits_{\gamma_2} F_x\,dx + F_y\,dy. \qquad (15)$$

Since $G = G_1 \cup G_2$, by virtue of the additivity of the double integral, the left-hand side of equality (15) is the double integral over the domain G, that is,

$$\iint\limits_{G_1} \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\, dy + \iint\limits_{G_2} \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\, dy = \iint\limits_{G} \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\, dy.$$

Let us rewrite either of the contour integrals from equality (15) in such a form:

$$\oint\limits_{\gamma_1} F_x\, dx + F_y\, dy = \int\limits_{\overset{\frown}{ABC}} F_x\, dx + F_y\, dy + \int\limits_{\overset{\frown}{CA}} F_x\, dx + F_y\, dy,$$

$$\oint\limits_{\gamma_2} F_x\, dx + F_y\, dy = \int\limits_{\overset{\frown}{CDA}} F_x\, dx + F_y\, dy + \int\limits_{\overset{\frown}{AC}} F_x\, dx + F_y\, dy.$$

From the equality $\int\limits_{\overset{\frown}{CA}} F_x\, dx + F_y\, dy = -\int\limits_{\overset{\frown}{AC}} F_x\, dx + F_y\, dy$ we conclude that the sum of the contour integrals is the integral over the boundary γ of the domain G, that is,

$$\oint\limits_{\gamma_1} F_x\, dx + F_y\, dy + \oint\limits_{\gamma_2} F_x\, dx + F_y\, dy = \int\limits_{\overset{\frown}{ABCDA}} F_x\, dx + F_y\, dy = \oint\limits_{\gamma} F_x\, dx + F_y\, dy.$$

In such a manner we make sure that formula (13) is true for the case when $G = G_1 \cup G_2$. The validity of Green's formula for the case $G = \bigcup\limits_{i=1}^{n} G_i$ is verified in a similar manner.

Green's formula also holds true in more general suppositions. This is shown in the following remarks.

Remark 1. Formula (13) remains true if the functions $F_x(x, y)$ and $F_y(x, y)$ are continuous in the closed domain $\bar{G}$, and their derivatives $\partial F_x/\partial y$, $\partial F_y/\partial x$ are continuous in an

open domain and there exists the integral

$$\iint\limits_G \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\, dy.$$

Remark 2. Green's formula is also true for the multiply connected domain G whose boundary consists of piecewise smooth contours (Fig. 14).

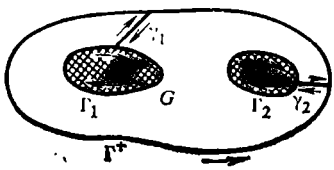


Fig. 14

To extend Green's theorem to a multiply connected domain, the latter should be turned into a simply connected domain, which is achieved as follows. Let Γ^+ be an outer contour, and let Γ_i^-, $i = 1, 2, \ldots, k$, be inner contours. We draw additional lines γ_i connecting the inner contours Γ_i^- and the outer contour Γ^+, then the domain G becomes simply connected. For the obtained domain with a continuous piecewise smooth contour

$$\Gamma = \Gamma^+ + \bigcup_{i=1}^{k} \Gamma_i^- + \bigcup_{i=1}^{k} \gamma_i^+ + \bigcup_{i=1}^{k} \gamma_i^-$$

we can write Green's formula. Taking into consideration that the integrals taken over the curves γ_i^+ and γ_i^- are mutually annihilated, we obtain Green's formula (16) for a multiply connected domain:

$$\iint\limits_G \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\, dy = \oint\limits_{\Gamma^+} F_x\, dx + F_y\, dy + \sum_{i=1}^{k} \oint\limits_{\Gamma_i^-} F_x\, dx + F_y\, dy. \qquad (16)$$

Sec. 1.4. FLUX OF A VECTOR FIELD

1. Surface Integrals of the Second Type. Let us introduce the notion of surface integrals of the second type with the aid of the surface integral of the first type $\iint\limits_\Sigma f(M)\, d\sigma$,

where Σ is a smooth surface with the continuous function $f(M)$ defined at each of its points.

Let Σ be a smooth oriented surface whose orientation is determined by the unit normal $\mathbf{n}^0$ and let $\mathbf{F}(M)$ be a vector function defined and continuous on Σ. The surface integral of the first type in the case when $f = \mathbf{F} \cdot \mathbf{n}^0$ is a scalar product of the vector $\mathbf{F}$ by the unit vector of the normal $\mathbf{n}^0$, that is, the integral

$$\iint_{\Sigma} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma \tag{1}$$

will be called the *surface integral of the second type* of the vector $\mathbf{F}$ through the surface Σ.

Sometimes, instead of expression (1), another designation is used for the surface integral of the second type. Let us denote by $d\sigma$ the vector whose length is numerically equal to the area of the element $d\sigma$ and whose direction coincides with that of the unit normal $\mathbf{n}^0$. Then expression (1) takes the form

$$\iint_{\Sigma} \mathbf{F} \cdot d\sigma = \iint_{\Sigma} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma.$$

The surface Σ on which the unit normal $\mathbf{n}^0$ is chosen will be denoted by the symbol Σ^+, while the surface Σ on which the normal $-\mathbf{n}^0$ is taken will be denoted by Σ^- ($\mathbf{n}^0$ and $-\mathbf{n}^0$ are the two orientations of the surface Σ). If the orientation of the surface is changed, then the surface integral will change its sign, that is,

$$\iint_{\Sigma^+} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma = -\iint_{\Sigma^-} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma.$$

Let us write integral (1) in extended form in the Cartesian coordinate system. The scalar product of the vectors

$$\mathbf{F} = F_x \mathbf{i} + F_y \mathbf{j} + F_z \mathbf{k} \quad \text{and}$$
$$\mathbf{n}^0 = \cos\alpha \, \mathbf{i} + \cos\beta \, \mathbf{j} + \cos\gamma \, \mathbf{k}$$

is expressed by the equality

$$\mathbf{F} \cdot \mathbf{n}^0 = F_x \cos\alpha + F_y \cos\beta + F_z \cos\gamma,$$

and integral (1) takes the form

$$\iint_{\Sigma} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma = \iint_{\Sigma} (F_x\cos\alpha + F_y\cos\beta + F_z\cos\gamma)\,d\sigma. \quad (2)$$

The quantities $d\sigma\cos\alpha$, $d\sigma\cos\beta$, $d\sigma\cos\gamma$ represent the projections of the surface element $d\sigma$ on the respective coordinate planes yOz, xOz, xOy, that is,[1]

$$d\sigma\cos\alpha = dy\,dz, \qquad d\sigma\cos\beta = dx\,dz, \qquad d\sigma\cos\gamma = dx\,dy.$$

We rewrite equality (2) as follows:

$$\iint_{\Sigma} F_x\,dy\,dz + F_y\,dx\,dz + F_z\,dx\,dy = \iint_{\Sigma} (F_x\cos\alpha + F_y\cos\beta + F_z\cos\gamma)\,d\sigma. \quad (3)$$

It follows from equality (3) that the surface integral of the second type (1) can be computed by projecting on the coordinate planes or with the aid of the surface integral of the first type on the right-hand side of equality (3).

2. **Flux of a Vector Field and Its Computation.** It is convenient to introduce the notion of the flux of a vector field by considering a particular case. Let $\mathbf{v} = \mathbf{v}(M)$ be a velocity field of a stationary fluid flux, and let Σ be a smooth oriented surface whose orientation is determined by the unit normal $\mathbf{n}^0 = \mathbf{n}^0(M)$. Let us compute the volume of the fluid Q flowing across the surface Σ per unit time.

Subdividing the surface Σ into n parts $\Sigma_1, \Sigma_2, \ldots, \Sigma_i, \ldots, \Sigma_n$, we find the quantity of fluid flowing across Σ_i per unit time. The relevant computation can be carried out in the following way. We assume that on Σ_i the rate of flux is constant and is equal to the value of the vector $\mathbf{v}(M_i)$ at some point $M_i \in \Sigma_i$. Then the quantity of the fluid flowing across the surface element Σ_i is equal to the volume of the cylindrical body with base Σ_i and generatrix $\mathbf{v}(M_i)$. The volume of such a body is approximately equal to the volume of an oblique cylinder with generatrix $\mathbf{v}(M_i)$ and base lying in the tangent plane passed at the point M_i

(Fig. 15), that is, it is equal to the product $\mathbf{v}(M_i)\cdot\mathbf{n}^0(M_i)\,\Delta\sigma_i$, where $\Delta\sigma_i$ is the area of the surface element Σ_i. Summing with respect to all surface elements Σ_i and passing to the limit, we find the volume of the flu l flowing across the surface Σ in the form of the surface integral of the second type:

$$Q=\iint_{\Sigma}\mathbf{v}\cdot\mathbf{n}^0\,d\sigma. \quad (4)$$

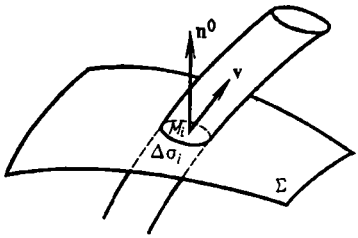


Fig. 15

Integral (4) is known as the *flux of the velocity vector* $\mathbf{v}$ across the surface Σ.

The notion of the flux is also introduced for an arbitrary vector field $\mathbf{F}=\mathbf{F}(M)$. Let Σ be a smooth oriented surface whose orientation is determined by the unit normal $\mathbf{n}^0=\mathbf{n}^0(M)$. The *flux of the vector field* $\mathbf{F}$ across the surface Σ is defined as the surface integral of the second type

$$\Pi=\iint_{\Sigma}\mathbf{F}\cdot\mathbf{n}^0\,d\sigma. \quad (5)$$

The methods of computation of the flux Π are based on the methods for evaluating surface integrals of the first and second types and require high skill.

Let us pay attention to the formula for computing the flux of the vector $\mathbf{F}$ across a surface represented implicitly: $\psi(x, y, z)=0$. In this case it may be regarded as the level surface $\psi(M)=c$ in a scalar field $\psi=\psi(x, y, z)$, the normal $\mathbf{n}^0$ to which is directed along the gradient towards increasing values of c and is written in the form

$$\mathbf{n}^0=\pm\frac{\operatorname{grad}\psi}{|\operatorname{grad}\psi|}.$$

From equality (5) we obtain the formula for computing the flux when the surface is represented implicitly: $\psi(x, y, z)=0$

$$\Pi=\pm\iint_{\Sigma}\frac{\mathbf{F}\cdot\operatorname{grad}\psi}{|\operatorname{grad}\psi|}\,d\sigma, \quad (6)$$

the sign being chosen depending on the orientation of the surface.

Example 1°. Find the flux of the vector field $\mathbf{F} = x\mathbf{i} + y\mathbf{j} - 3z\mathbf{k}$ across the outer surface of the paraboloid $z = x^2 + y^2$ cut away by the plane $z = 4$.

Let us compute the gradient of the scalar field $\psi = x^2 + y^2 - z$:

$$\operatorname{grad} \psi = 2x\mathbf{i} + 2y\mathbf{j} - \mathbf{k},$$

it is directed towards increasing values of $\psi = c$. In this case the direction of the gradient coincides with that of the outer normal, therefore for the outer normal we choose the plus sign; then

$$\mathbf{n}^0 = \frac{2x\mathbf{i} + 2y\mathbf{j} - \mathbf{k}}{\sqrt{4x^2 + 4y^2 + 1}}.$$

Substituting this expression into equality (5), we obtain the expression for computing the flux with the aid of the surface integral:

$$\Pi = \iint_{\Sigma} \frac{2x^2 + 2y^2 + 3z}{\sqrt{4x^2 + 4y^2 + 1}}\, d\sigma.$$

Using the expression for the surface element

$$d\sigma = \sqrt{1 + z_x'^2 + z_y'^2}\, dx\, dy = \sqrt{1 + 4x^2 + 4y^2}\, dx\, dy,$$

we reduce the evaluation of the surface integral to that of the double integral

$$\Pi = 5 \iint_{G} (x^2 + y^2)\, dx\, dy$$

(the range of integration G being the circle $x^2 + y^2 \leqslant 4$).

It seems to be more appropriate to perform the computation in the polar coordinate system $x = \rho \cos \varphi$, $y = \rho \sin \varphi$, $dx\, dy = \rho\, d\rho\, d\varphi$. Substituting these values into the integral, we obtain

$$\Pi = 5 \int_0^{2\pi} d\varphi \int_0^2 \rho^3\, d\rho = 40\pi.$$

Example 2°. Find the flux of the vector field $\mathbf{F} = z\mathbf{i} + \mathbf{j} - x\mathbf{k}$ across the outer part of the sphere $x^2 + y^2 + z^2 = 1$ situated in the first octant.

The outer normal to the sphere is directed along the radius, therefore on the sphere of the unit radius $r = 1$ the unit normal is representable in the form

$$\mathbf{n}^0 = \frac{\mathbf{r}}{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}.$$

We then find the scalar product $\mathbf{F} \cdot \mathbf{n}^0 = xz + y - xz = y$. To compute the flux $\Pi = \iint_\Sigma y \, d\sigma$ it is advisable to pass over to the spherical coordinate system. On the sphere of the unit radius we have the relationships $y = \sin\theta \sin\varphi$ and $d\sigma = \sin\theta \, d\theta \, d\varphi$, therefore

$$\Pi = \int_0^{\pi/2} \sin\varphi \, d\varphi \int_0^{\pi/2} \sin^2\theta \, d\theta = \frac{\pi}{2}.$$

3. Ostrogradsky-Gauss Formula. Green's formula proved in Sec. 1.3 established a relation between the line integral taken over a contour and the double integral over the domain bounded by this contour. It turns out that there exists an analogous relation between the surface integral over a closed surface and the triple integral over the domain bounded by this surface. To find out this relation, let us agree on the following terminology.

We shall not give the definition of a closed surface, since it is too complicated, but will confine ourselves to its intuitive understanding. By a *closed surface* we shall understand a surface which is the boundary of a certain bounded spatial body. We can show that any piecewise smooth closed surface is oriented. In this case one of the orientations consists of the unit normals directed from the surface into the domain Ω, that is, of the so-called inner normals, and the other orientation of the unit normals directed from the surface out of the domain Ω, that is, of the outer normals.

The spatial domain Ω will be called simple with respect to the z-axis if its projection G_{xy} on the plane xOy is a rectifiable (squarable) domain, and the boundary consists of

the surfaces $z = z_1(x, y)$ and $z = z_2(x, y)$ and a part of the cylinder whose base is the domain G_{xy}. Here we suppose that the functions $z_1(x, y)$ and $z_2(x, y)$ are continuous and satisfy the inequality $z_1(x, y) \leqslant z_2(x, y)$ on G_{xy} (Fig. 16). The domains simple with respect to the x- and y-axes are defined in the same manner by considering the equations of the surface $x = x(y, z)$ and $y = y(x, z)$ and the bases of a cylinder G_{yz} and G_{xz} lying in the planes yOz and xOz, respectively.

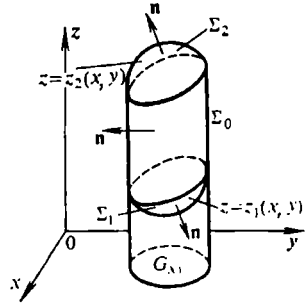


Fig. 16

The domain Ω is said to be *simple* if it can be split into a finite number of subdomains, simple with respect to the three coordinate axes simultaneously.

Theorem 1. *If the vector function* $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ *together with the partial derivatives* $\partial F_x/\partial x$, $\partial F_y/\partial y$, $\partial F_z/\partial z$ *are continuous in the closure of the simple domain* Ω, *then there holds true the Ostrogradsky-Gauss formula*

$$\iiint_{\Omega} \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}\right) dv = \oiint_{\Sigma} F_x\, dy\, dz + F_y\, dx\, dz + F_z\, dx\, dy, \quad (7)$$

where the surface integrals are taken over the outer side of the surface Σ.

Let us prove formula (7) for a domain which is simple with respect to the z-axis (see Fig. 16). The boundary Σ of this domain consists of three smooth orientable surfaces:

$$\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_0,$$

where Σ_1 and Σ_2 are the surfaces defined in the domain G_{xy} by the equations $z = z_1(x, y)$ and $z = z_2(x, y)$ and Σ_0 is a cylindrical surface with generatrix parallel to the z-axis. The positive orientation of the surfaces Σ_1 and Σ_2 is their outer sides Σ_1^+ and Σ_2^+.

Let us transform the triple integral over the domain Ω of the function $\partial F_z/\partial z$ to the double integral over the projection G_{xy}:

$$\iiint\limits_{\Omega} \frac{\partial F_z}{\partial z}\,dx\,dy\,dz = \iint\limits_{G_{xy}} \left(\int\limits_{z_1(x,\,y)}^{z_2(x,\,y)} \frac{\partial F_z}{\partial z}\,dz \right) dx\,dy$$

$$= \iint\limits_{G_{xy}} [F_z(x,\ y,\ z_2(x,\ y)) - F_z(x,\ y,\ z_1(x,\ y))]\,dx\,dy.$$

We then express the double integrals in terms of the surface integrals over Σ_1^+ and Σ_2^+, taking into consideration their respective orientation:

$$\iint\limits_{G_{xy}} F_z(x,\ y,\ z_2(x,\ y))\,dx\,dy = \iint\limits_{\Sigma_2^+} F_z\,dx\,dy,$$

$$\iint\limits_{G_{xy}} F_z(x,\ y,\ z_1(x,\ y))\,dx\,dy = -\iint\limits_{\Sigma_1^+} F_z\,dx\,dy;$$

then the triple integral will be represented in the form of the sum of surface integrals:

$$\iiint\limits_{\Omega} \frac{\partial F_z}{\partial z}\,dv = \iint\limits_{\Sigma_2^+} F_z\,dx\,dy + \iint\limits_{\Sigma_1^+} F_z\,dx\,dy.$$

The cylindrical surface Σ_0 has the generatrix parallel to the z-axis, therefore the normal $\mathbf{n}^0$ to the surface Σ_0 makes the angle $\gamma = \pi/2$ with the z-axis and $\cos\gamma = 0$. Consequently, the surface integral

$$\iint\limits_{\Sigma_0} F_z\,dx\,dy = \iint\limits_{\Sigma_0} F_z \cos\gamma\,d\sigma = 0.$$

Thus, the triple integral over the domain Ω can be expressed in terms of the surface integral over the boundary Σ:

$$\iiint\limits_{\Omega} \frac{\partial F_z}{\partial z}\,dv = \iint\limits_{\Sigma_2^+} F_z\,dx\,dy + \iint\limits_{\Sigma_1^+} F_z\,dx\,dy$$

$$+ \iint\limits_{\Sigma_0} F_z\,dx\,dy = \oiint\limits_{\Sigma} F_z\,dx\,dy.$$

For the domains simple with respect to the y- and x-axes, the same as in the case of the domain simple with respect to the z-axis, the following formulas are proved:

$$\iiint_{\Omega} \frac{\partial F_y}{\partial y}\, dv = \oiint_{\Sigma} F_y\, dx\, dz \quad \text{and}$$

$$\iiint_{\Omega} \frac{\partial F_x}{\partial x}\, dv = \oiint_{\Sigma} F_x\, dy\, dz.$$

If the domain Ω is simple simultaneously with respect to all the coordinate axes, then, adding together the last three formulas, we obtain the Ostrogradsky-Gauss formula (7) for this domain.

The Ostrogradsky-Gauss formula will be true for a simple domain as well. Indeed, let us partition Ω into simple with respect to all axes subdomains Ω_ν, $\nu = 1, 2, \ldots, k$, and writing for each of them formula (7), add together the results obtained. Then, by virtue of the additivity of the integral, on the left-hand side we obtain the integral over the domain Ω. Further, taking into consideration that the outer normals to the inner parts of the boundaries of the subdomains Ω_ν are directed in opposite sides, we obtain that the sum of the surface integrals over these parts of the boundaries of the subdomains Ω_ν will be equal to zero. Hence, the right-hand side of the obtained sum will retain only the integrals over those parts of the boundaries of the subdomains Ω_ν which jointly make up the boundary of the domain Ω. It is advisable to partition a domain into subdomains by planes parallel to the coordinate planes.

Remark 1. The Ostrogradsky-Gauss formula also holds for more general conditions: if the function $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ is continuous in the closed domain $\bar{\Omega}$ and the derivatives $\partial F_x/\partial x$, $\partial F_y/\partial y$, $\partial F_z/\partial z$ are continuous in the open domain Ω, and the integral $\iiint_{\Omega} \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}\right) dv$ exists.

Remark 2. The Ostrogradsky-Gauss formula is also valid for the multiply connected domain Ω whose boundary con

sists of a finite number of piecewise smooth surfaces. In this case the surface Σ represents the union of the outer surface Σ^+ and inner surfaces Σ_i^-, $i = 1, 2, \ldots, n$, i.e

$$\Sigma = \Sigma^+ + \bigcup_{i=1}^{n} \Sigma_i^-.$$

Sec. 1.5. DIVERGENCE

1. The Notion of Divergence. Consider the flux of the velocity field of a fluid once again. If the vectors $\mathbf{v}$ and $\mathbf{n}^0$ form an acute angle, then the quantity $\mathbf{v} \cdot \mathbf{n}^0$ is positive, and if $\mathbf{v}$ and $\mathbf{n}^0$ form an obtuse angle, then this quantity is negative. Therefore the flux Q determined by formula (4) derived in Sec. 1.4, generally speaking, represents the excess of the fluid flowing in the direction of the positive normal $\mathbf{n}^0$ but not the absolute quantity of the fluid flown across the surface Σ independently of the direction of flow.

The magnitude of the flux Q of the field across a closed surface Σ can be regarded as the difference between the quantity of the fluid entering the domain Ω and flowing out of it. If the flux is positive, then this means that the outflow of the fluid exceeds its inflow, and, vice versa, if the flux Q is negative, then the inflow exceeds the outflow. If the flux Q is equal to zero, then this can mean either the absence of sources and sinks or, at least, the presence of such a distribution of sources and sinks that their total power is zero.

The magnitude of the flux of a vector across a closed surface is a global characteristic of the field in the domain Ω and enables us to judge of the presence of sources and sinks in this domain, but very approximately. In connection with this, it is advisable to introduce a local characteristic of distribution of sources and sinks. Such a characteristic is divergence (flux density at a point).

Let us give the definition of the divergence at a point. We surround a fixed point M_0 of a vector field by an arbitrary closed smooth surface Σ which bounds the domain Ω_0. The average value of the flux is the ratio of the flux of the vector $\mathbf{F}$ across the surface Σ to the magnitude of the volume $v =$

$= v(\Omega_0)$ enclosed inside Σ:

$$\frac{1}{v(\Omega_0)} \oiint_{\Sigma} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma. \tag{1}$$

In hydrodynamic interpretation of the flux of a vector, the quantity (1) is referred to as the *average power* of sources and sinks per unit volume. The limit of relationship (1) is the *flux density* at a point. (Physically, the divergence measures the "flux per unit volume" across the surface of a small element situated at a certain point.)

If there exists the limit of relationship (1), as the surface Σ contracts to a point M_0, and this limit is independent of the form of the surface Σ, then it will be called the *divergence of the field* $\mathbf{F}$ at the point M_0. We will write it in the following way:

$$\operatorname{div} \mathbf{F}(M_0) = \lim_{\substack{\Sigma \to M_0 \\ (v \to 0)}} \frac{\oiint_{\Sigma} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma}{v(\Omega_0)}, \tag{2}$$

where the symbol $\Sigma \to M_0$ means that the surface Σ contracts to the point M_0, here, of course, the volume $v \to 0$. Such a definition of divergence is independent of the choice of a coordinate system.

2. Computing the Divergence in Cartesian Coordinates. Let us find the formula for the divergence in the Cartesian coordinate system.

Theorem 1. *If the vector field* $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ *and the partial derivatives* $\partial F_x/\partial x$, $\partial F_y/\partial y$, $\partial F_z/\partial z$ *are defined and continuous in the domain* Ω, *then at any point* $M(x, y, z)$ *of this domain there exists the divergence* $\operatorname{div} \mathbf{F}(M)$ *and the following formula is valid:*

$$\operatorname{div} \mathbf{F}(M) = \frac{\partial F_x(M)}{\partial x} + \frac{\partial F_y(M)}{\partial y} + \frac{\partial F_z(M)}{\partial z}. \tag{3}$$

Applying the Ostrogradsky-Gauss formula (7) from the preceding section, we can write the equality

$$\oiint_{\Sigma_0} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma = \iiint_{\Omega_0} \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} \right) dv,$$

where $\Omega_0 \subset \Omega$ and Σ_0 is the surface bounding the domain Ω_0; further, by the mean-value theorem, we have:

$$\oiint_{\Sigma_0} \mathbf{F}\cdot\mathbf{n}^0\, d\sigma = \left(\frac{\partial F_x(M_1)}{\partial x} + \frac{\partial F_y(M_1)}{\partial y} + \frac{\partial F_z(M_1)}{\partial z}\right) v(\Omega_0),$$

$$M_1 \in \Omega_0.$$

Substituting the obtained expression into (2) and taking into consideration the continuity of the partial derivatives, we make sure that equality (3) holds true:

$$\operatorname{div}\mathbf{F}(M) = \lim_{\substack{\Sigma_0 \to M \\ (M_1 \to M)}} \left(\frac{\partial F_x(M_1)}{\partial x} + \frac{\partial F_y(M_1)}{\partial y} + \frac{\partial F_z(M_1)}{\partial z}\right)$$

$$= \frac{\partial F_x(M)}{\partial x} + \frac{\partial F_y(M)}{\partial y} + \frac{\partial F_z(M)}{\partial z}.$$

3. Properties of the Divergence. Let us note some simplest properties of the divergence.

(a) *From the linearity of partial derivatives there follows the property of linearity of the divergence*:

$$\operatorname{div}(c_1\mathbf{F}_1 + c_2\mathbf{F}_2) = c_1 \operatorname{div}\mathbf{F}_1 + c_2 \operatorname{div}\mathbf{F}_2,\quad c_1, c_2\text{—const.}$$

(b) *Let $f = f(M)$ be a scalar function, and let $\mathbf{F} = \mathbf{F}(M)$ be a vector function. Then the divergence of the product $f\mathbf{F}$ is representable in the form of the sum*:

$$\operatorname{div}(f\mathbf{F}) = f \operatorname{div}\mathbf{F} + \mathbf{F}\cdot\operatorname{grad} f. \tag{4}$$

Indeed, the following equality is valid:

$$\operatorname{div}(f\mathbf{F}) = \frac{\partial(fF_x)}{\partial x} + \frac{\partial(fF_y)}{\partial y} + \frac{\partial(fF_z)}{\partial z}$$

$$= f \operatorname{div}\mathbf{F} + \left(F_x\frac{\partial f}{\partial x} + F_y\frac{\partial f}{\partial y} + F_z\frac{\partial f}{\partial z}\right).$$

But the expression in the parentheses is the scalar product $\mathbf{F}\cdot\operatorname{grad} f$.

In particular, if $\mathbf{F} = \mathbf{a}$ is a constant vector, then $\operatorname{div}\mathbf{a} = 0$ and formula (4) takes the form

$$\operatorname{div}(f\mathbf{a}) = \mathbf{a}\operatorname{grad} f,$$

and if $f = c =$ constant, then grad $c = 0$ and formula (4) takes the form

$$\operatorname{div}(c\mathbf{F}) = c \operatorname{div} \mathbf{F}.$$

Example 1°. Find div F for the field determined by the expression $\mathbf{F} = f(r)\cdot\mathbf{r}$, which is usually called *central-symmetric.*

Using formula (4), we write the divergence in the form

$$\operatorname{div} \mathbf{F} = \operatorname{div}(f(r)\,\mathbf{r}) = f(r)\operatorname{div}\mathbf{r} + \mathbf{r}\operatorname{grad} f(r).$$

We then simplify the obtained expression. Applying formula (3), we compute the divergence of the radius vector:

$$\operatorname{div}\mathbf{r} = \operatorname{div}(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) = 3.$$

By formula (9), derived in Sec. 1.2, we have

$$\operatorname{grad} f(r) = f'(r)\frac{\mathbf{r}}{r}.$$

Substituting the values of div r and grad $f(r)$ into the expression for div F, we get the formula for computing the divergence of the central-symmetric field:

$$\operatorname{div}(f(r)\,\mathbf{r}) = 3f(r) + rf'(r). \tag{5}$$

Example 2°. A body rotates about the z-axis anticlockwise with a constant angular velocity $\boldsymbol{\omega} = \omega\mathbf{k}$. Find the divergence of the velocity vector v.

The velocity v is the vector product $\boldsymbol{\omega} \times \mathbf{r}$; computing this product, we obtain

$$\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & \omega \\ x & y & z \end{vmatrix} = -\omega y\mathbf{i} + \omega x\mathbf{j}.$$

By formula (3), we find: $\operatorname{div}\mathbf{v} = \operatorname{div}(-\omega y\mathbf{i} + \omega x\mathbf{j}) = 0$.

4. The Ostrogradsky-Gauss Formula in Vector Form. Using the notions of divergence and flux, let us give a vector interpretation of the Ostrogradsky-Gauss formula. The surface integral of the second type in this formula represents the flux, that is,

$$\iint_{\Sigma} P_x\,dy\,dz + P_y\,dx\,dz + P_z\,dx\,dy = \iint_{\Sigma} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma.$$

Using the formula

$$\operatorname{div}\mathbf{F}=\frac{\partial F_x}{\partial x}+\frac{\partial F_y}{\partial y}+\frac{\partial F_z}{\partial z},$$

we rewrite the triple integral in formula (7) from the preceding section in the form

$$\iiint\limits_{\Omega}\left(\frac{\partial F_x}{\partial x}+\frac{\partial F_y}{\partial y}+\frac{\partial F_z}{\partial z}\right)d v=\iiint\limits_{\Omega}\operatorname{div}\mathbf{F}\,d v,$$

therefore the Ostrogradsky-Gauss formula can be written in vector form:

$$\oiint\limits_{\Sigma}\mathbf{F}\cdot\mathbf{n}^0\,d\sigma=\iiint\limits_{\Omega}\operatorname{div}\mathbf{F}\,d v. \tag{6}$$

This notation means that the flux of the vector $\mathbf{F}$ across the outer side of the closed surface Σ is equal to the integral of the divergence of the field $\mathbf{F}$ taken over the domain Ω bounded by this surface.

The Ostrogradsky-Gauss formula in form (6) is independent of the choice of coordinate system. It can be applied to compute a flux through a closed surface.

Example 3°. Compute the flux of the vector $\mathbf{F}=(x-1)^3\,\mathbf{i}+(y+2)^3\,\mathbf{j}+(z-2)^3\,\mathbf{k}$ across the outer side of the sphere $(x-1)^2+(y+2)^2+(z-2)^2=R^2$.

Applying formula (6), we reduce the computation of the desired flux to the evaluation of a triple integral:

$$\Pi=\oiint\limits_{\Sigma}\mathbf{F}\cdot\mathbf{n}^0\,d\sigma=\iiint\limits_{\Omega}\operatorname{div}\left[(x-1)^3\,\mathbf{i}+(y+2)^3\,\mathbf{j}+(z-2)^3\,\mathbf{k}\right]d v=3\iiint\limits_{\Omega}\left[(x-1)^2+(y+2)^2+(z-2)^2\right]d v.$$

The computation of the last integral is convenient to be performed in the spherical coordinate system. Let us set $x-1=r\sin\theta\cos\varphi$, $y+2=r\sin\theta\sin\varphi$, $z-2=r\cos\theta$, then $dv=r^2\sin\theta\,dr\,d\theta\,d\varphi$. Substituting these

expressions into the integral, we obtain

$$\Pi = 3 \iiint_{\Omega} r^4 \sin\theta \, dr \, d\theta \, d\varphi$$

$$= 3 \int_0^{2\pi} d\varphi \int_0^{\pi} \sin\theta \, d\theta \int_0^{R} r^4 dr = \frac{12}{5} \pi R^5.$$

Example 4°. Compute the flux of the vector $\mathbf{F} = \mathbf{r}$ across a closed surface.

Since $\operatorname{div} \mathbf{r} = 3$, by formula (6), we obtain

$$\Pi = \oiint_{\Sigma} \mathbf{r} \cdot \mathbf{n}^0 \, d\sigma = 3 \iiint_{\Omega} dv = 3V(\Omega).$$

Thus, the flux of the radius vector $\mathbf{r}$ across the closed surface Σ is equal to the triple volume of the body enclosed inside this surface.

5. Computing the Divergence in Orthogonal Curvilinear Coordinates. We can show that the divergence of the vecto-field $\mathbf{F} = F^1\mathbf{e}_1 + F^2\mathbf{e}_2 + F^3\mathbf{e}_3$ given in the curvilinear coorr dinate system is computed by the formula

$$\operatorname{div} \mathbf{F} = \frac{1}{H_1 H_2 H_3} \left(\frac{\partial (F^1 H_2 H_3)}{\partial q^1} + \frac{\partial (F^2 H_3 H_1)}{\partial q^2} + \frac{\partial (F^3 H_1 H_2)}{\partial q^3} \right). \quad (7)$$

Using formulas (16) and (18) from Sec. 1.1, from (7) we find the divergence expressed in the cylindrical coordinate system:

$$\operatorname{div} \mathbf{F} = \frac{1}{\rho} \frac{\partial (\rho F_\rho)}{\partial \rho} + \frac{1}{\rho} \frac{\partial F_\varphi}{\partial \varphi} + \frac{\partial F_z}{\partial z} \quad (8)$$

and in the spherical coordinate system:

$$\operatorname{div} \mathbf{F} = \frac{1}{r^2} \frac{\partial (r^2 F_r)}{\partial r} + \frac{1}{r \sin\theta} \frac{\partial (\sin\theta F_\theta)}{\partial \theta} + \frac{1}{r \sin\theta} \frac{\partial F_\varphi}{\partial \varphi}. \quad (9)$$

Example 5°. Compute the divergence of the field $\mathbf{F} = \rho^2 \boldsymbol{\rho}$ in the cylindrical coordinate system.

We write the vector field F in the form $\mathbf{F} = \rho^3 \mathbf{e}_\rho$, and then, by formula (8), we find

$$\operatorname{div} \mathbf{F} = \frac{1}{\rho}(\rho^4)' = 4\rho^2 = 4(x^2 + y^2).$$

Example 6°. Compute the divergence of the field $\mathbf{F} = \mathbf{r}/r^4$ in the spherical coordinate system.

Bearing in mind that $\mathbf{F} = (1/r^3)\,\mathbf{e}_r$ and using formula (9), we have

$$\operatorname{div} \mathbf{F} = \frac{1}{r^2}\left(\frac{1}{r}\right)' = -\frac{1}{r^3}.$$

The same result can also be obtained with the aid of formula (5).

Sec. 1.6. THE CURL OF A VECTOR FIELD

1. Circulation of a Vector Field. The work of the vector field F round a closed curve γ is of a specific interest. In this case the work of the vector field F is called the *circulation* of the vector F round the curve γ and is denoted as

$$C = \oint_\gamma \mathbf{F} \cdot d\mathbf{r}. \tag{1}$$

Let us clarify the physical meaning of circulation. We shall interpret the field F as the velocity field $\mathbf{v} = \mathbf{v}(M)$ of a flowing fluid. Let us place in this field a small wheel with blades fitted round the rim γ of this wheel (Fig. 17). The particles of the fluid acting on these blades will generate a large number of torques whose total effect will set the wheel in rotary motion about its axis. The rotary effect of the velocity field of the fluid will be characterized at each point M (Fig. 18) by the projection of the vector $\mathbf{v}(M)$ on the tangent line τ^0 to the circle γ, that is, by the scalar product $\mathbf{v} \cdot \tau^0$. The summation of the rotary actions of the fluid round the entire rim of the wheel leads to the notion of circulation (1) of the vector $\mathbf{F} = \mathbf{v}$. Here, the absolute value of circulation will determine the angular velocity of the rotating wheel, and the sign of the circulation will show in what side the wheel rotates relative to the direction chosen on it.

Using the terminology of the velocity field of a flowing fluid, we may say that the circulation of an arbitrary field $\mathbf{F}(M)$ determines its "rotary capacity" in a given direction and characterizes the vorticity of the field in this direction.

Under the sign of the integral in formula (1) we see a scalar product, therefore the circulation depends not only on the absolute values of the vectors but also on the angles between the vector field and the tangents to the curve γ.

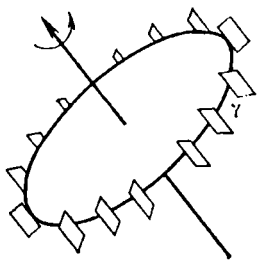


Fig. 17

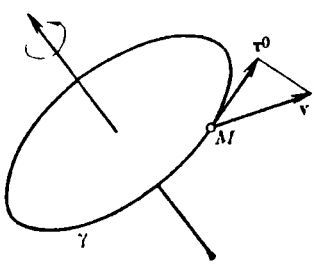


Fig. 18

The less the angle, the greater the circulation, and, hence, the greater the vorticity of the field in this direction.

Example 1°. Compute the circulation of the plane vector field $\mathbf{F} = y^2\mathbf{i} + x\mathbf{j}$ round the curve $x = 3\cos t$, $y = \sin t$ traversed clockwise.

The given curve is an ellipse. Since the curve is traversed clockwise, t varies from 2π to 0. Consequently, the circulation is computed in the following way:

$$C = \oint_{\gamma} \mathbf{F}\cdot d\mathbf{r} = \oint_{\gamma} y^2\,dx + x\,dy = \int_{2\pi}^{0} (-3\sin^3 t + 3\cos^2 t)\,dt$$

$$= \frac{3}{2}\int_{2\pi}^{0} (1-\cos 2t)\,dt = -3\pi.$$

2. The Notion of Curl. This notion is closely connected with the notion of circulation. Determining the rotational capacity of a field about a given direction, the circulation characterizes the vorticity of a vector field along the entire

contour. The local characteristic of the field associated with vorticity is the curl.

Let us first consider a plane vector field **F** and a certain contour γ surrounding a chosen point M_0. The area of the region enclosed by the contour γ will be denoted by S. Then the relationship

$$\frac{1}{S}\oint_{\gamma}\mathbf{F}\cdot d\mathbf{r} \tag{2}$$

gives the average circulation density of the vector **F** on the area S. The circulation density at the point M_0 is characterized by the limit of expression (2), provided the contour γ is contracted to the point M_0; then the area S surrounded by the contour γ tends to zero. Thus, if the limit

$$\lim_{\substack{\gamma\to M_0\\(S\to 0)}}\frac{1}{S}\oint_{\gamma}\mathbf{F}\cdot d\mathbf{r} \tag{3}$$

exists, then it gives the vorticity of the field at the point M_0.

If the vector field **F** is spatial, then we may speak of the vorticity of the field in a certain direction **n**. Let us pass through the point M_0 a plane π perpendicular to the chosen direction **n**, and let us consider in it a contour γ enclosing the point M_0 (Fig. 19). Then, finding limit (3), we obtain the vorticity of the field in the direction **n**. This limit underlies the definition of the curl of the vector field **F**.

The *curl* of the vector field **F** at the point M_0, denoted as curl **F**, is defined as the vector whose projection on each direction **n** is equal to the limit of the ratio of the circulation of the vector field round the contour γ of the plane domain G, which is perpendicular to this direction, to the surface area S of this domain, as the area tends to zero and the domain

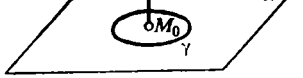


Fig. 19

itself contracts to the point M_0, that is,

$$\operatorname{pr}_{\mathbf{n}} \operatorname{curl} \mathbf{F} = (\operatorname{curl} \mathbf{F})_{\mathbf{n}} = \lim_{\substack{\gamma \to M_0 \\ (S \to 0)}} \frac{1}{S} \oint_{\gamma} \mathbf{F} \cdot d\mathbf{r}, \tag{4}$$

where γ is the contour lying in the plane perpendicular to the vector $\mathbf{n}$, and S is the surface area of the domain bounded by this contour.

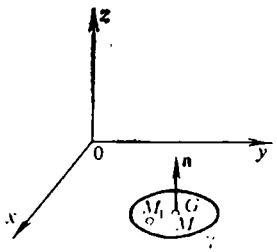


Fig. 20

The given definition of curl is independent of the coordinate system chosen.

3. Computing the Curl in Cartesian Coordinates. Let us find the formula for computing the curl of a vector in the Cartesian coordinate system.

Theorem 1. *If a continuously differentiable vector field* $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ *is given in a certain domain* Ω, *then at every point* $M \in \Omega$ *there exists* curl $\mathbf{F}$ *which is computed by the formula*

$$\operatorname{curl} \mathbf{F} = \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right)\mathbf{i} + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right)\mathbf{j} + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)\mathbf{k}. \tag{5}$$

Let us compute, for instance, the projection of the curl on the z-axis. Let γ be a contour lying in the (x, y)-plane, and let S be the surface area of the domain G bounded by this contour (Fig. 20). Using Green's formula from Sec. 1.3, we can write

$$\oint_{\gamma} \mathbf{F} \cdot d\mathbf{r} = \oint_{\gamma} F_x\,dx + F_y\,dy = \iint_{G} \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right) dx\,dy.$$

Applying the mean-value theorem to the last integral, we find

$$\oint_{\gamma} \mathbf{F} \cdot d\mathbf{r} = \left(\frac{\partial F_y(M_1)}{\partial x} - \frac{\partial F_x(M_1)}{\partial y}\right) S,$$

where M_1 is some point of the domain G. Substituting this integral into (4), we obtain the expression for the projection of curl F on the z-axis:

$$(\operatorname{curl} \mathbf{F})_z = \lim_{\substack{M_1 \to M \\ (\gamma \to M)}} \left(\frac{\partial F_y(M_1)}{\partial x} - \frac{\partial F_x(M_1)}{\partial y} \right) = \frac{\partial F_y(M)}{\partial x} - \frac{\partial F_x(M)}{\partial y}.$$

In the same way, we find the projections of the curl on the x- and y-axes:

$$(\operatorname{curl} \mathbf{F})_x = \frac{\partial F_z(M)}{\partial y} - \frac{\partial F_y(M)}{\partial z},$$

$$(\operatorname{curl} \mathbf{F})_y = \frac{\partial F_x(M)}{\partial z} - \frac{\partial F_z(M)}{\partial x}.$$

4. Hamiltonian Operator. The fundamental notions of gradient, divergence, and curl introduced in vector analysis can be represented in a convenient form by means of a special symbolic vector operator ∇ (introduced by the Irish mathematician W. R. Hamilton) called *nabla*, *del*, or the *Hamiltonian operator*. It is defined by the formula

$$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k},$$

where $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ symbolize the differentiation with respect to x, y, and z, or in other words:

$$\nabla_x = \frac{\partial}{\partial x}, \quad \nabla_y = \frac{\partial}{\partial y}, \quad \nabla_z = \frac{\partial}{\partial z}.$$

The vector ∇ itself has no real meaning, but it becomes meaningful in combination with scalar or vector functions. Thus, the basic notions of the field theory: grad f, div F, and curl F can be expressed with the aid of the operator ∇. To this end, let us define the "product" of the operator ∇ by a scalar function $f(x, y, z)$ as the product of the appropriate projections ∇_x, ∇_y, and ∇_z by this function, the "products" $\nabla_x f$, $\nabla_y f$, and $\nabla_z f$ being understood as taking the appropriate partial derivatives of the function $f(x, y, z)$, that is,

$$\nabla_x f = \frac{\partial f}{\partial x}, \quad \nabla_y f = \frac{\partial f}{\partial y}, \quad \nabla_z f = \frac{\partial f}{\partial z}.$$

Consequently, the product ∇f is defined by the vector

$$\nabla f = \frac{\partial f}{\partial x}\,\mathbf{i} + \frac{\partial f}{\partial y}\,\mathbf{j} + \frac{\partial f}{\partial z}\,\mathbf{k},$$

but this is just grad f. Thus, the product of the symbolic operator ∇ by a scalar function f results in the gradient of this scalar function, that is, $\nabla f = \operatorname{grad} f$.

Now, using the rule for a scalar product of vectors, let us determine the scalar product of the vector ∇ by a vector function $\mathbf{F} = \mathbf{F}(x, y, z) = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$:

$$\nabla\cdot\mathbf{F} = \left(\frac{\partial}{\partial x}\,\mathbf{i} + \frac{\partial}{\partial y}\,\mathbf{j} + \frac{\partial}{\partial z}\,\mathbf{k}\right)\cdot(F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k})$$

$$= \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}.$$

But the last expression is just the divergence of the vector $\mathbf{F} = \mathbf{F}(x, y, z)$, i.e. $\nabla\cdot\mathbf{F} = \operatorname{div}\mathbf{F}$.

Finally, let us define the vector product of the vector ∇ by the vector $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ according to the rule for an ordinary product of vectors with the aid of the equality

$$\nabla\times\mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & F_z \end{vmatrix}$$

$$= \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right)\mathbf{i} + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right)\mathbf{j} + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)\mathbf{k}.$$

The right-hand side represents the curl of the vector function $\mathbf{F}$, therefore $\nabla\times\mathbf{F} = \operatorname{curl}\mathbf{F}$.

Proceeding from the above reasoning, we conclude that, with the aid of the Hamiltonian operator, the fundamental notions of the field theory, obtained by means of the operation of differentiation, can be written in the form of various products used in vector algebra.

Let us enumerate a number of properties of the operator ∇ whose proof follows immediately from the above given definitions of the appropriate products.

If α, β are constants, f_p are scalar functions, and $\mathbf{F}_p$, $p = 1, 2$ are vector functions, then

A. $\nabla(\alpha f_1 + \beta f_2) = \alpha\nabla f_1 + \beta\nabla f_2$;

B. $\nabla\cdot(\alpha\mathbf{F}_1 + \beta\mathbf{F}_2) = \alpha\nabla\cdot\mathbf{F}_1 + \beta\nabla\cdot\mathbf{F}_2$;

C. $\nabla\times(\alpha\mathbf{F}_1 + \beta\mathbf{F}_2) = \alpha\nabla\times\mathbf{F}_1 + \beta\nabla\times\mathbf{F}_2$;

D. $\nabla(f_1 f_2) = f_1\nabla f_2 + f_2\nabla f_1$;

E. $\nabla\cdot(f\mathbf{F}) = f\nabla\cdot\mathbf{F} + \mathbf{F}\cdot\nabla f$;

F. $\nabla\times(f\mathbf{F}) = f\nabla\times\mathbf{F} + \nabla f\times\mathbf{F}$;

G. $\nabla\cdot(\mathbf{F}_1\times\mathbf{F}_2) = \mathbf{F}_2\cdot\nabla\times\mathbf{F}_1 - \mathbf{F}_1\cdot\nabla\times\mathbf{F}_2$.

The properties D to G express certain properties of the above-introduced differential operators grad f, div $\mathbf{F}$, and curl $\mathbf{F}$. These properties can also be written as follows:

D_1. grad $(f_1 f_2) = f_1$ grad $f_2 + f_2$ grad f_1;

E_1. div $(f\mathbf{F}) = f$ div $\mathbf{F} + \mathbf{F}\cdot$grad f;

F_1. curl $(f\mathbf{F}) = f$ curl $\mathbf{F} +$ (grad f) $\times\mathbf{F}$;

G_1. div $(\mathbf{F}_1\times\mathbf{F}_2) = \mathbf{F}_2\cdot$curl $\mathbf{F}_1 - \mathbf{F}_1\cdot$curl $\mathbf{F}_2$.

Example 2°. Find the curl of the velocity field of a solid rotating about a fixed point with an instantaneous angular velocity $\boldsymbol{\omega} = \omega_x\mathbf{i} + \omega_y\mathbf{j} + \omega_z\mathbf{k}$.

As is known, the velocity of a solid is determined by the formula

$$\mathbf{v} = \boldsymbol{\omega}\times\mathbf{r} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \omega_x & \omega_y & \omega_z \\ x & y & z \end{vmatrix}$$

$$= (z\omega_y - y\omega_z)\,\mathbf{i} + (x\omega_z - z\omega_x)\,\mathbf{j} + (y\omega_x - x\omega_y)\,\mathbf{k}.$$

Hence we find curl $\mathbf{v}$:

$$\operatorname{curl}\mathbf{v} = \nabla\times\mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ z\omega_y - y\omega_z & x\omega_z - z\omega_x & y\omega_x - x\omega_y \end{vmatrix}$$

$$= 2\omega_x\mathbf{i} + 2\omega_y\mathbf{j} + 2\omega_z\mathbf{k} = 2\boldsymbol{\omega}.$$

Thus, curl $\mathbf{F}$ characterizing the "rotary component" of the velocity field is equal to twice the rotation velocity.

Example 3°. Find curl grad f.

Using the operator ∇, we can write: curl grad $f = \nabla \times (\nabla f)$. But the vector ∇f is collinear with the vector ∇, and therefore their vector product equals zero, that is,

$$\text{curl grad } f = 0. \tag{6}$$

Example 4°. Show that the vorticity of a field reaches the greatest value in the direction of the curl.

The vorticity of the field F in the direction n is equal to the projection of the curl on this direction, that is,

$$(\text{curl F})_n = |\text{ curl F }| \cos(\widehat{\mathbf{n}, \text{curl F}}).$$

Hence, it is seen that the field F has the greatest vorticity in the case when $\cos(\widehat{\mathbf{n}, \text{curl F}}) = 1$, and this means that the direction of the normal n coincides with the direction of curl F, the maximal vorticity being equal to $|\text{ curl F }|$.

5. Curl in an Orthogonal Curvilinear Coordinate System. It is possible to show that the curl of the vector field $\mathbf{F} = F^1\mathbf{e}_1 + F^2\mathbf{e}_2 + F^3\mathbf{e}_3$ given in curvilinear coordinates is computed by the formula

$$\begin{aligned}\text{curl F} = &\frac{1}{H_2H_3}\left(\frac{\partial(H_3F^3)}{\partial q^2} - \frac{\partial(H_2F^2)}{\partial q^3}\right)\mathbf{e}_1 \\ &+ \frac{1}{H_3H_1}\left(\frac{\partial(H_1F^1)}{\partial q^3} - \frac{\partial(H_3F^3)}{\partial q^1}\right)\mathbf{e}_2 \\ &+ \frac{1}{H_1H_2}\left(\frac{\partial(H_2F^2)}{\partial q^1} - \frac{\partial(H_1F^1)}{\partial q^2}\right)\mathbf{e}_3.\end{aligned}$$

This formula is convenient to be written in symbolic form

$$\text{curl F} = \frac{1}{H_1H_2H_3}\begin{vmatrix} H_1\mathbf{e}_1 & H_2\mathbf{e}_2 & H_3\mathbf{e}_3 \\ \dfrac{\partial}{\partial q^1} & \dfrac{\partial}{\partial q^2} & \dfrac{\partial}{\partial q^3} \\ H_1F^1 & H_2F^2 & H_3F^3 \end{vmatrix}. \tag{7}$$

Using formulas (16) and (18) derived in Sec. 1.1, we obtain the expression for the curl in cylindrical coordinates:

$$\begin{aligned}\text{curl F} = &\left(\frac{1}{\rho}\frac{\partial F_z}{\partial \varphi} - \frac{\partial F_\varphi}{\partial z}\right)\mathbf{e}_\rho + \left(\frac{\partial F_\rho}{\partial z} - \frac{\partial F_z}{\partial \rho}\right)\mathbf{e}_\varphi \\ &+ \left(\frac{1}{\rho}\frac{\partial(\rho F_\varphi)}{\partial \rho} - \frac{1}{\rho}\frac{\partial F_\rho}{\partial \varphi}\right)\mathbf{e}_z = \frac{1}{\rho}\begin{vmatrix} \mathbf{e}_\rho & \rho\mathbf{e}_\varphi & \mathbf{e}_z \\ \dfrac{\partial}{\partial \rho} & \dfrac{\partial}{\partial \varphi} & \dfrac{\partial}{\partial z} \\ F_\rho & \rho F_\varphi & F_z \end{vmatrix}\end{aligned} \tag{8}$$

and in spherical coordinates:

$$\operatorname{curl}\mathbf{F}=\frac{1}{r\sin\theta}\left(\frac{\partial(F_\varphi\sin\theta)}{\partial\theta}-\frac{\partial F_\theta}{\partial\varphi}\right)\mathbf{e}_r$$
$$+\left(\frac{1}{r\sin\theta}\frac{\partial F_r}{\partial\varphi}-\frac{1}{r}\frac{\partial(rF_\varphi)}{\partial r}\right)\mathbf{e}_\theta+\left(\frac{1}{r}\frac{\partial(rF_\theta)}{\partial r}-\frac{1}{r}\frac{\partial F_r}{\partial\theta}\right)\mathbf{e}_\varphi$$
$$=\frac{1}{r^2\sin\theta}\begin{vmatrix}\mathbf{e}_r & r\mathbf{e}_\theta & r\sin\theta\mathbf{e}_\varphi\\ \frac{\partial}{\partial r} & \frac{\partial}{\partial\theta} & \frac{\partial}{\partial\varphi}\\ F_r & rF_\theta & r\sin\theta F_\varphi\end{vmatrix}. \quad (9)$$

Example 5°. Find the curl of the vector field $\mathbf{F}$ given in spherical coordinates: $\mathbf{F}=\frac{\cos\theta}{r^2}\mathbf{e}_r+\frac{\sin\theta}{r^2}\mathbf{e}_\theta$.

By formula (9), we find

$$\operatorname{curl}\mathbf{F}=\frac{1}{r^2\sin\theta}\begin{vmatrix}\mathbf{e}_r & r\mathbf{e}_\theta & r\sin\theta\mathbf{e}_\varphi\\ \frac{\partial}{\partial r} & \frac{\partial}{\partial\theta} & \frac{\partial}{\partial\varphi}\\ \frac{\cos\theta}{r^2} & \frac{\sin\theta}{r^2} & 0\end{vmatrix}$$
$$=\frac{\mathbf{e}_\varphi}{r}\left(\frac{\sin\theta}{r^2}-\frac{\sin\theta}{r^2}\right)=0.$$

6. Stokes' Formula. Stokes' formula is Green's formula generalized for a three-dimensional case. It relates the circulation of a vector field with the curl flux across the surface spanned on this contour. We will say that the surface Σ is spanned on a piecewise smooth contour Γ if there exists a piecewise smooth oriented surface Σ lying in the domain Ω and having the contour Γ as its boundary (Fig. 21).

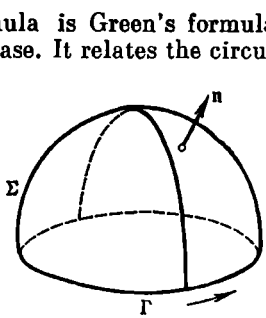


Fig. 21

A three-dimensional domain Ω is called *simply connected* if any closed surface in it bounds a region completely lying in Ω. Otherwise the domain is called *multiply connected*. The entire space, the ball, the parallelepiped are examples

of simply connected domains. The domain enclosed between two concentric spheres is an example of a doubly connected domain.

A three-dimensional domain Ω is the so-called *superficially simply connected* if on any contour Γ lying in Ω it is possible to span a surface Σ also entirely lying in Ω.

The notion of a superficially simply connected domain is not equivalent to the notion of a simply connected domain.

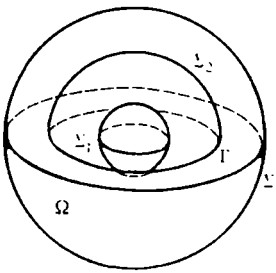


Fig. 22

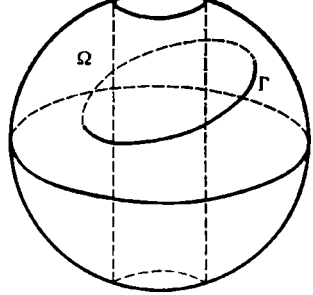


Fig. 23

The entire space, ball, parallelepiped are examples of superficially simply connected domains.

As it has already been indicated, the domain enclosed between two concentric spheres Σ and Σ_1 is not simply connected, but at the same time this domain is superficially simply connected (Fig. 22). On any contour Γ, including one enclosing the inner sphere Σ_1, we can span the surface Σ_2 entirely belonging to the domain Ω.

A superficially not simply connected domain may be exemplified by a ball through which there passes a cylindrical tunnel (Fig. 23). On the contours Γ enveloping this cylinder, it is impossible to span a surface belonging to the domain Ω.

Consider an oriented surface Σ bounded by one or several contours Γ. Here, the direction of the normal to the surface Σ is made consistent with the direction of traversing the contour Γ according to the following rule: the direction in which the contour Γ is traced will be regarded as *positive*

(*consistent with the orientation of* Σ) if an observer situated on the surface so that the direction of the normal coincides with the direction from the observer's legs to his head traces the contour Γ keeping the surface Σ on his left. The opposite direction will be regarded as *negative* (see Fig. 21).

Stokes' Theorem. *Let* Ω *be a superficially simply connected domain,* Γ *be a piecewise smooth contour in* Ω, *and* Σ *be a piecewise smooth surface spanned on the contour* Γ *and lying in* Ω. *Let in the domain* Ω *there be given a vector field* $\mathbf{F} = \mathbf{F}(M)$ *such that* $\mathbf{F}(M)$ *and* curl $\mathbf{F}(M)$ *are continuous in the domain* Ω. *Then the circulation of the field* $\mathbf{F}(M)$ *round the contour* Γ *is equal to the flux of the curl* curl $\mathbf{F}$ *across the surface* Σ, *that is,*

$$\oint_{\Gamma} \mathbf{F} \cdot d\mathbf{r} = \int\int_{\Sigma} \operatorname{curl} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma, \tag{10}$$

the direction of the contour Γ *and the orientation of the surface,* Σ *being coordinated.*

Let us partition the surface Σ into k parts Σ_ν bounded with respective contour Γ_ν, $\nu = 1, 2, \ldots, k$. Consider the νth element Σ_ν of the surface Σ. Let us take an arbitrary point $M_\nu \subset \Sigma_\nu$, and draw through this point a normal $\mathbf{n}_\nu$ and a tangent plane π_ν to the surface Σ. Now we denote the projection of the contour Γ_ν on the plane π_ν by γ_ν, the area of the surface Σ_ν by $\Delta\sigma_\nu$, and the area of the projection of the surface Σ_ν on the plane π_ν by ΔS_ν (Fig. 24). The definition of curl (4) implies the equality

$$\oint_{\gamma_\nu} \mathbf{F} \cdot d\mathbf{r} = [\operatorname{curl} \mathbf{F}(M_\nu)]_{\mathbf{n}_\nu} \Delta S_\nu + o(\Delta S_\nu).$$

For sufficiently small subsurfaces this equality will also hold true for the contour Γ_ν of the surface Σ_ν:

$$\oint_{\Gamma_\nu} \mathbf{F} \cdot d\mathbf{r} = [\operatorname{curl} \mathbf{F}(M_0)]_{\mathbf{n}_\nu} \Delta\sigma_\nu + o(\Delta\sigma_\nu), \quad \nu = 1, 2, \ldots, k.$$

Summing together the obtained equalities with respect to all ν, we find the relationship

$$\sum_{\nu=1}^{k} \oint_{\Gamma_\nu} \mathbf{F} \cdot d\mathbf{r} = \sum_{\nu=1}^{k} \{[\operatorname{curl} \mathbf{F}(M_0)]_{\mathbf{n}_\nu} \Delta\sigma_\nu + o(\Delta\sigma_\nu)\}. \tag{11}$$

Let us transform the left-hand side of equality (11). When joining two neighbouring sections of the surfaces Σ_ν and Σ_s according to the rule for agreement of the directions of the contour and the surface, the common part of their boundary will be traversed in opposite directions (Fig. 25). Hence it follows that, when summing together two circulations, the

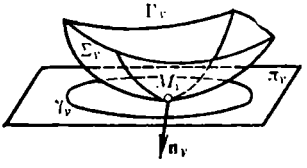


Fig. 24

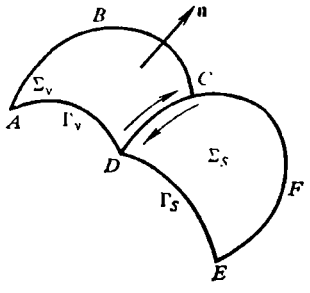


Fig. 25

integrals taken over the common part of the boundary are annihilated, and there remains the circulation along the contour bounding the union of the surfaces $\Sigma_\nu \cup \Sigma_s$

$$\oint_{\Gamma_\nu} \mathbf{F}\cdot d\mathbf{r} + \oint_{\Gamma_s} \mathbf{F}\cdot d\mathbf{r} = \underbrace{\oint}_{ADCBA} \mathbf{F}\cdot d\mathbf{r} + \underbrace{\oint}_{DEFCD} \mathbf{F}\cdot d\mathbf{r} = \underbrace{\oint}_{AEFBA} \mathbf{F}\cdot d\mathbf{r}.$$

Summing together the contour integrals with respect to all ν, we obtain the integral over the common contour, that is,

$$\sum_{\nu=1}^{k} \oint_{\Gamma_\nu} \mathbf{F}\cdot d\mathbf{r} = \oint_{\Gamma} \mathbf{F}\cdot d\mathbf{r}.$$

Equality (11) takes the form

$$\oint_{\Gamma} \mathbf{F}\cdot d\mathbf{r} = \sum_{\nu=1}^{k} \{[\operatorname{curl} \mathbf{F}(M_\nu)]_{\mathbf{n}_\nu} \Delta\sigma_\nu + o(\Delta\sigma_\nu)\}. \qquad (12)$$

The sum $\sum_{\nu=1}^{k} [\operatorname{curl} \mathbf{F}(M_\nu)]_{\mathbf{n}_\nu} \Delta\sigma_\nu$ is integral for the surface integral $\iint_\Sigma \operatorname{curl} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma$. Passing to the limit in equality (12), we obtain equality (10).

Stokes' formula (10) in Cartesian coordinates has the form

$$\oint_\Gamma F_x \, dx + F_y \, dy + F_z \, dz = \iint_\Sigma \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right) dy \, dz + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \right) dz \, dx + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) dx \, dy. \quad (13)$$

Remark 1. It follows from Stokes' formula that if two surfaces Σ_1 and Σ_2 are spanned on the contour Γ, then the fluxes of curl $\mathbf{F}$ across these surfaces are equal.

Remark 2. Let $\mathbf{F}$ be a plane field, say, it is defined in the xy-plane by the formula $\mathbf{F} = F_x(x, y)\,\mathbf{i} + F_y(x, y)\,\mathbf{j}$. Then for curl $\mathbf{F}$ we obtain the representation

$$\operatorname{curl} \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & 0 \end{vmatrix} = \mathbf{k} \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right)$$

and Stokes' formula (13) takes the form

$$\oint_\Gamma F_x \, dx + F_y \, dy = \iint_\Sigma \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) dx \, dy.$$

In particular, if a plane domain G lying inside the contour Γ is taken for the surface Σ, then in this case Stokes' formula turns into Green's formula (13) given in Sec. 1.3. Therefore, Green's formula may be given the following vector interpretation: the circulation of a plane vector field is equal to the flux of its curl across the domain lying inside the contour.

Given in this chapter, the notions of gradient, curl, divergence and also Green's, Stokes', and the Ostrogradsky-Gauss formulas are widely used in physics and technology in many applications when real physical fields are considered.

CHAPTER 2

Special Kinds of Vector Fields

Sec. 2.1. POTENTIAL VECTOR FIELD

1. Notion and Properties of a Potential Field. A vector field $\mathbf{F}(M)$ is said to be *potential* in a domain Ω if there exists a scalar field $f(M)$ such that for all points of this domain $\mathbf{F}(M)$ is the field of the gradient of this scalar field $f(M)$, that is,

$$\mathbf{F}(M) = \operatorname{grad} f(M). \tag{1}$$

In this case, the scalar field $f(M)$ is called the *potential* of the vector field $\mathbf{F}(M)$.

A potential field is one of the simplest vector fields, since it is completely defined by one scalar function $f(M)$, i.e. by the potential, whereas an arbitrary vector field $\mathbf{F}(M)$ is defined by three scalar functions, i.e. by the projections F_x, F_y, F_z. Here and elsewhere, we will assume that the field $\mathbf{F}$ is continuously differentiable in the domain under consideration.

Theorem 1. *If the field* $\mathbf{F}$ *is potential, then its potential is determined by this field uniquely to within an arbitrary constant term.*

Let us suppose that the field $\mathbf{F}$ has two potentials f_1 and f_2, i.e. $\mathbf{F} = \operatorname{grad} f_1$ and $\mathbf{F} = \operatorname{grad} f_2$. Then $\operatorname{grad}(f_1 - f_2) \equiv 0$. Hence it follows that $f_2 = f_1 + c$. (See Property (d) of gradient in Sec. 1.2.)

Theorem 2. *If the field* $\mathbf{F}$ *is potential in a domain* Ω, *then the work is independent of the shape of the path and the potential* $f(M)$ *is determined with an accuracy up to an arbi-*

trary constant using the line integral of the second kind

$$f(M)=\int\limits_{(M_0)}^{(M)} F_x\,dx+F_y\,dy+F_z\,dz \qquad (2)$$

taken over an arbitrary curve $\gamma\in\Omega$ *joining the points* M_0 *and* M. *Here,* $M_0(x_0, y_0, z_0)$ *is some fixed point from* Ω, *and* $M(x, y, z)\in\Omega$ *is an arbitrary (variable) point.*

The work A of the field F done along some path γ joining the points M_0 and M is computed by the formula

$$A=\int\limits_{M_0M}\mathbf{F}\cdot d\mathbf{r}=\int\limits_{M_0M}F_x\,dx+F_y\,dy+F_z\,dz.$$

Since the field F is potential, there exists a potential f such that $\mathbf{F}=\operatorname{grad} f$. In this case the scalar product

$$\mathbf{F}\cdot d\mathbf{r}=\operatorname{grad} f\cdot d\mathbf{r}=\frac{\partial f}{\partial x}dx+\frac{\partial f}{\partial y}dy+\frac{\partial f}{\partial z}dz=df$$

is the total differential of the potential f. Therefore, the work A is equal to the difference of the potentials at the terminal and initial points:

$$A=\int\limits_{M_0M}df=f(M)-f(M_0).$$

It is seen from this equality that the work depends not on the shape of the path but only on the positions of the initial and terminal points. The integral independent of the shape of the path is used to be written in the form $\int\limits_{(M_0)}^{(M)}\mathbf{F}\,d\mathbf{r}$.

Comparing two expressions for the work A, we obtain for the potential f formula (2) with an accuracy to the constant $f(M_0)$.

The problem of finding the potential $f(M)$ of the vector field $\mathbf{F}(M)$ is closely connected with the problem of restoring a function of three variables from its total differential.

Theorem 3. *Let the functions* $P(M)$, $Q(M)$, $R(M)$ *be continuous and have continuous partial derivatives in the domain* Ω. *For the expression*

$$P(M)\,dx+Q(M)\,dy+R(M)\,dz \qquad (3)$$

to be the total differential of some function $\varphi(M)$, *it is necessary and sufficient that the vector field*

$$\Phi(M) = P(M)\,\mathbf{i} + Q(M)\,\mathbf{j} + R(M)\,\mathbf{k} \qquad (4)$$

be potential.

Necessity. Let (3) be the total differential of the function $\varphi(M)$, that is,

$$d\varphi = P\,dx + Q\,dy + R\,dz.$$

This means that $P = \partial\varphi/\partial x$, $Q = \partial\varphi/\partial y$, and $R = \partial\varphi/\partial z$, therefore

$$\operatorname{grad}\varphi = \frac{\partial\varphi}{\partial x}\,\mathbf{i} + \frac{\partial\varphi}{\partial y}\,\mathbf{j} + \frac{\partial\varphi}{\partial z}\,\mathbf{k} = P\mathbf{i} + Q\mathbf{j} + R\mathbf{k} = \Phi.$$

Consequently, the field Φ is potential.

Sufficiency. Let the vector field $\Phi(M)$ defined by (4) be potential, that is, there exists a scalar function $\varphi(M)$ such that

$$\operatorname{grad}\varphi = \Phi(M) = P(M)\,\mathbf{i} + Q(M)\,\mathbf{j} + R(M)\,\mathbf{k}.$$

Hence, by virtue of the definition of gradient, we have the following equalities:

$$P = \frac{\partial\varphi}{\partial x}, \qquad Q = \frac{\partial\varphi}{\partial y}, \qquad R = \frac{\partial\varphi}{\partial z}.$$

And this means that

$$d\varphi = \frac{\partial\varphi}{\partial x}\,dx + \frac{\partial\varphi}{\partial y}\,dy + \frac{\partial\varphi}{\partial z}\,dz = P\,dx + Q\,dy + R\,dz,$$

that is, expression (3) is the total differential.

From Theorems 2 and 3 there directly follows the method of restoring the function $\varphi(x, y, z)$ from its total differential $P\,dx + Q\,dy + R\,dz$. To determine φ it is sufficient to use formula (2), that is, to evaluate, to within an arbitrary constant, the line integral

$$\varphi(x, y, z) = \int_{(x_0, y_0, z_0)}^{(x, y, z)} P\,dx + Q\,dy + R\,dz + c.$$

Let us also note that if it is known that the expression $P\,dx + Q\,dy + R\,dz$ is the total differential of the function

$\varphi(x, y, z)$, then for computing the line integral

$$\int_{(x_0, y_0, z_0)}^{(x, y, z)} P\,dx + Q\,dy + R\,dz$$

we have the following analogue of the Newton-Leibniz formula:

$$\int_{(x_0, y_0, z_0)}^{(x, y, z)} P\,dx + Q\,dy + R\,dz = \varphi(x, y, z) - \varphi(x_0, y_0, z_0).$$

2. Field Potentiality Conditions. Naturally, there arises a question concerning the conditions under which a given vector field $\mathbf{F}(M)$ will be potential. The answer to this question is given by Theorem 4. But first we shall prove the following lemma.

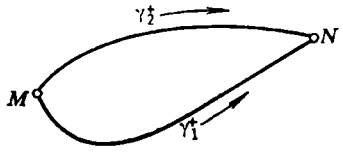


Fig. 26

Lemma. *In order for the work to be independent of the shape of the path, it is necessary and sufficient that the circulation round any closed contour be equal to zero.*

Necessity. Let the work be independent of the shape of the path. In the domain Ω, we take an arbitrary closed contour Γ which can be regarded as the union of the curves γ_1^+ and γ_2^- joining the points M and N, i.e. $\Gamma = \gamma_1^+ \cup \gamma_2^-$ (Fig. 26). By the hypothesis, the work is independent of the shape of the path, i.e.

$$\int_{\gamma_1^+} \mathbf{F}\cdot d\mathbf{r} = \int_{\gamma_2^+} \mathbf{F}\cdot d\mathbf{r} = -\int_{\gamma_2^-} \mathbf{F}\cdot d\mathbf{r},$$

therefore for circulation we have the representation

$$\oint_{\Gamma} \mathbf{F}\cdot d\mathbf{r} = \int_{\gamma_1^+} \mathbf{F}\cdot d\mathbf{r} - \int_{\gamma_2^-} \mathbf{F}\cdot d\mathbf{r} = 0.$$

Thus, the circulation round an arbitrary contour equals zero.

Sufficiency. Let $\oint_\Gamma \mathbf{F}\cdot d\mathbf{r} = 0$, where Γ is an arbitrary contour in the domain Ω. Consider two arbitrary paths connecting the points M and N (see Fig. 26). The union of the curves $\gamma_1^+ \cup \gamma_2^-$ forms a closed contour Γ, therefore $\oint_\Gamma \mathbf{F}\cdot d\mathbf{r} = 0$. By the properties of integrals, we have the equalities

$$\oint_\Gamma \mathbf{F}\cdot d\mathbf{r} = \int_{\gamma_1^+} \mathbf{F}\cdot d\mathbf{r} + \int_{\gamma_2^-} \mathbf{F}\cdot d\mathbf{r} = \int_{\gamma_1^+} \mathbf{F}\cdot d\mathbf{r} - \int_{\gamma_2^+} \mathbf{F}\cdot d\mathbf{r}.$$

Hence we conclude that the work is independent of the shape of the path, since $\int_{\gamma_1^+} \mathbf{F}\cdot d\mathbf{r} = \int_{\gamma_2^+} \mathbf{F}\cdot d\mathbf{r}$.

Theorem 4. *In order for a continuous differentiable vector field* $\mathbf{F}(M)$ *to be potential in a superficially simply connected domain* Ω, *it is necessary and sufficient that it be irrotational, i.e. that* $\operatorname{curl}\mathbf{F}(M) = 0$ *for all points* $M \subset \Omega$.

Necessity. Let the field $\mathbf{F}$ be potential, that is, $\mathbf{F} = \operatorname{grad} f = \nabla f$. Bearing in mind Example 3° from Sec. 1.6, we make sure that $\operatorname{curl}\mathbf{F} = \nabla \times \nabla f = 0$, i.e. the field $\mathbf{F}$ is irrotational.

Sufficiency. Let the field $\mathbf{F}(M)$ be irrotational, i.e. $\operatorname{curl}\mathbf{F}(M) = 0 \ \forall\, M \subset \Omega$. Since the domain Ω is superficially simply connected, we conclude from Stokes' theorem that for any closed contour $\Gamma \subset \Omega$

$$\oint_\Gamma \mathbf{F}\cdot d\mathbf{r} = 0.$$

The lemma implies that the integral

$$\int_\gamma \mathbf{F}\cdot d\mathbf{r} = \int_\gamma F_x\,dx + F_y\,dy + F_z\,dz \qquad (5)$$

is independent of the shape of the curve $\gamma \subset \Omega$ connecting the points $M_0(x_0, y_0, z_0)$ and $M(x, y, z)$ and depends only on the points M_0 and M. This means that if M_0 is a fixed point of the domain Ω, then integral (5) is the function of the point M. Let us denote this function by $f(M)$ and show

that grad $f = \mathbf{F}$; thereby it will be proved that the field $\mathbf{F}$ is potential. Thus, integral (5) is the function $f(x, y, z)$, that is,

$$f(x, y, z) = \int_{(x_0, y_0, z_0)}^{(x, y, z)} F_x\,dx + F_y\,dy + F_z\,dz. \qquad (6)$$

Integral (6) is independent of the path of integration, therefore the difference $f(x + \Delta x, y, z) - f(x, y, z)$ can be represented in the form

$$f(x+\Delta x, y, z) - f(x, y, z) = \int_{(x, y, z)}^{(x+\Delta x, y, z)} F_x\,dx + F_y\,dy + F_z\,dz,$$

where the line segment, taken as the path of integration, joins the points (x, y, z) and $(x + \Delta x, y, z)$ and is parallel to the x-axis. Consequently, $dy = 0$ and $dz = 0$, and the preceding equality takes the form

$$f(x+\Delta x, y, z) - f(x, y, z) = \int_{(x, y, z)}^{(x+\Delta x, y, z)} F_x\,dx = F_x(x + \theta\,\Delta x, y, z)\,\Delta x, \quad [0 < \theta < 1.$$

Writing the last equality, we used the mean-value theorem. Taking into account that F_x is continuous, we conclude that at the point M there exists the partial derivative $\partial f/\partial x$ and the following equality holds:

$$\frac{\partial f}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x, y, z) - f(x, y, z)}{\Delta x} = F_x(x, y, z).$$

Making use of the same method, we find $\partial f/\partial y = F_y$ and $\partial f/\partial z = F_z$, and this just means that $f(x, y, z)$ is a potential and its gradient coincides with the field $\mathbf{F}$:

$$\operatorname{grad} f = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}.$$

Remark. When proving Theorem 4, we supposed that the domain Ω was superficially simply connected. But if Ω is multiply connected, then the condition $\operatorname{curl} \mathbf{F} = 0$, generally speaking, is not sufficient for the field $\mathbf{F}(M)$

to be potential. The necessary and sufficient condition of the potentiality of a field is established by the following theorem.

Theorem 5. *The necessary and sufficient condition of the potentiality of a field in both simply connected and multiply connected domains is the equality to zero of the circulation of the field round any contour.*

Necessity. Let the field **F** be potential, then, by Theorem 2, the work done by the field is independent of the shape of the path, and, by virtue of the lemma, the circulation is equal to zero.

Sufficiency. Let the circulation round any closed contour be equal to zero, then, by the same lemma, the work A, representable in the form of the integral

$$A = \int_{\gamma} \mathbf{F} \cdot d\mathbf{r} = \int_{\gamma} F_x\,dx + F_y\,dy + F_z\,dz,$$

is independent of the shape of the path. Reasoning in the same way as in proving Theorem 4, we come to a conclusion that the field is potential.

If the domain Ω is multiply connected, then in it there can exist contours on which it is impossible to span a surface wholly situated in the domain of the field. In this case, from the condition curl $\mathbf{F} = 0$ it does not follow, generally speaking, that the circulation of the field round any contour is equal to zero. This means that there can exist contours round which the circulation is different from zero: $\oint_{\gamma} \mathbf{F} \cdot d\mathbf{r} \neq 0$, and the field **F** will not be potential. The following example illustrates this assertion.

Example 1°. Show that the intensity of the magnetic field of an infinite straight conductor is not a potential field. Compute the circulation round the contour enveloping the z-axis.

As is known, the intensity of the magnetic field of the infinite straight conductor situated along the z-axis is given by the expression $\mathbf{H} = \frac{2I}{x^2+y^2}(-y\mathbf{i} + x\mathbf{j})$ (see Example 1 given in Sec. 1.3) defined everywhere except for the points of the z-axis. Therefore, the vector field is given in a doubly

connected domain Ω representing the whole three-dimensional space with the z-axis cut away.

Let us show that curl $\mathbf{H} = 0$ for all points of the domain Ω:

$$\operatorname{curl} \mathbf{H} = 2I\nabla x \left(\frac{-y\mathbf{i}+x\mathbf{j}}{x^2+y^2}\right) = 2I \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ -\frac{y}{x^2+y^2} & \frac{x}{x^2+y^2} & 0 \end{vmatrix}$$

$$= 2I\mathbf{k}\left[\frac{\partial}{\partial x}\left(\frac{x}{x^2+y^2}\right)+\frac{\partial}{\partial y}\left(\frac{y}{x^2+y^2}\right)\right]$$

$$= 2I\mathbf{k}\,\frac{x^2+y^2-2x^2+x^2+y^2-2y^2}{(x^2+y^2)^2} = 0.$$

Hence it follows that round any closed contour γ not enveloping the z-axis, the circulation is equal to zero, that is,

$$C = \oint_{\gamma} \mathbf{H}\cdot d\mathbf{r} = 0.$$

And if the contour Γ envelops the z-axis once, then the circulation $C = \oint_{\Gamma} \mathbf{H}\cdot d\mathbf{r}$ is different from zero, but is independent of its form.

In particular, if we take for the contour Γ a circle in the xy-plane with centre at the origin and of radius R: $\mathbf{r}(t) = R\cos t\,\mathbf{i} + R\sin t\,\mathbf{j}$, then after evaluation, we obtain

$$\oint_{\Gamma} \mathbf{H}\cdot d\mathbf{r} = 2I \oint_{\Gamma} \frac{-y\,dx+x\,dy}{x^2+y^2} = 2I\int_0^{2\pi} \frac{R^2\,dt}{R^2} = 4\pi I.$$

Consequently, round any contour developing the z-axis once the circulation attains the value $\oint_{\Gamma} \mathbf{H}\cdot d\mathbf{r} = 4\pi I$.

If the contour Γ envelops the z-axis k times, then the circulation will be equal to $\oint_{\Gamma} \mathbf{H}\cdot d\mathbf{r} = 4\pi kI$.

Theorem 5 implies that the intensity of the magnetic field is not a potential field in the doubly connected domain Ω.

3. Methods of Finding the Potential. The potential f of potential field $\mathbf{F}$ is found by the formula

$$f(M) = \int_{M_0}^{M} \mathbf{F} \cdot d\mathbf{r}, \tag{7}$$

where M_0 is a fixed and M an arbitrary point in the domain Ω.

Let us first see how the potential is found in Cartesian coordinates. Let $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$. Then the expression $\mathbf{F} \cdot d\mathbf{r} = F_x\,dx + F_y\,dy + F_z\,dz$ is the total differential of the potential f of the field, that is, $F_x\,dx + F_y\,dy + F_z\,dz = df$. Therefore the potential f is found with the aid of formula (2). Practically, the most convenient path of integration in the line integral (2) is the broken line M_0KNM whose segments are parallel to the coordinate axes (provided these segments belong to the domain Ω) (Fig. 27). Thus, we may write:

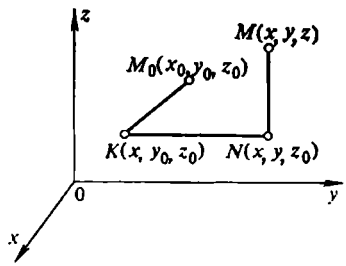


Fig. 27

$$f(M) = \int_{M_0K} \mathbf{F} \cdot d\mathbf{r} + \int_{KN} \mathbf{F} \cdot d\mathbf{r} + \int_{NM} \mathbf{F} \cdot d\mathbf{r} + c.$$

Integrating over the straight line M_0K, we have: $dy = dz = 0$, therefore $d\mathbf{r} = \mathbf{i}\,dx$, and the integration over the segment M_0K leads to the integral

$$\int_{M_0K} \mathbf{F} \cdot d\mathbf{r} = \int_{M_0K} F_x\,dx = \int_{x_0}^{x} F_x(x, y_0, z_0)\,dx.$$

Proceeding in the same manner, we reduce the integrals over the segments KN and NM to the respective definite

integrals:

$$\int_{KN} \mathbf{F}\cdot d\mathbf{r} = \int_{y_0}^{y} F_y(x, y, z_0)\, dy,$$

$$\int_{NM} \mathbf{F}\cdot d\mathbf{r} = \int_{z_0}^{z} F_z(x, y, z)\, dz.$$

Consequently, the potential f, with an accuracy to a constant, is determined by the equality

$$f(M) = \int_{x_0}^{x} F_x(x, y_0, z_0)\, dx + \int_{y_0}^{y} F_y(x, y, z_0)\, dy + \int_{z_0}^{z} F_z(x, y, z)\, dz. \quad (8)$$

Example 2°. Check to see that the field $\mathbf{F} = (3yz + x^2)\,\mathbf{i} + (2y^2 + 3xz)\,\mathbf{j} + (z^2 + 3xy)\,\mathbf{k}$ is potential, and find its potential.

Let us establish that the field is irrotational. Indeed, for an arbitrary point M we have:

$$\operatorname{curl}\mathbf{F}(M) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 3yz + x^2 & 2y^2 + 3xz & z^2 + 3xy \end{vmatrix} = \mathbf{i}(3x - 3x) + \mathbf{j}(3y - 3y) + \mathbf{k}(3z - 3z) = 0.$$

Hence, the field $\mathbf{F}$ is potential in the whole space. To compute the potential f, we take the point M_0 for the origin; then, applying formula (8), we obtain

$$f(M) = \int_0^x x^2\, dx + \int_0^y 2y^2\, dy + \int_0^z (z^2 + 3yx)\, dz + c = \frac{1}{3}(x^3 + 2y^3 + z^3) + 3xyz + c.$$

Example 3°. Let $\mathbf{F} = -\frac{\gamma m}{r^3}\mathbf{r}$ be a gravitational field which represents the force of attraction of the unit mass

placed at the point M (Fig. 28) by the mass m situated at the origin. The force is defined at all points except at the origin and forms a vector field—the gravity field of the point mass m. Show that the field F is potential in the whole space, except for the origin, and find its potential.

The relevant computations are more convenient to be carried out in spherical coordinates; we write the vector field in the form

$$\mathbf{F} = -\frac{\gamma m}{r^2}\frac{\mathbf{r}}{r} = -\frac{\gamma m}{r^2}\mathbf{e}_r.$$

Then the projections of the vector F on the coordinate axes have the form

$$F_r = -\frac{\gamma m}{r^2}, \quad F_\theta = 0, \quad F_\varphi = 0.$$

Applying formula (9) from Sec. 1.6, we find that curl F = $= 0$ for any $r \neq 0$. Bearing in mind that the whole space with a punctured point is superficially simply connected, we conclude that the field F is potential in this space.

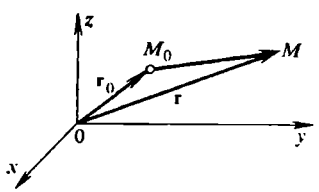


Fig. 28

It is advisable to take the vector $\overrightarrow{M_0M} = \mathbf{r} - \mathbf{r}_0$ for the path of integration in integral (7) (see Fig. 28). To compute the integrand in integral (7), we find the differential $d(\mathbf{r} - \mathbf{r}_0) = d\mathbf{r} = \mathbf{e}_r\, dr$ and the scalar product

$$\mathbf{F} \cdot d(\mathbf{r} - \mathbf{r}_0) = \mathbf{F} \cdot d\mathbf{r} = -\frac{\gamma m}{r^2}\mathbf{e}_r \cdot \mathbf{e}_r\, dr = -\frac{\gamma m}{r^2}dr.$$

Thus, the potential of the gravitational field is determined to within a constant by the expression

$$f = -\gamma m \int_{r_0}^{r} \frac{dr}{r^2} = \frac{\gamma m}{r} + c.$$

The obtained result may also be given an electrostatic interpretation. If a positive charge e is placed at the origin,

and the unit charge at the point M, then, according to Coulomb's law, the voltage of the electrostatic field at the point M is given by the equality $\mathbf{E} = (e/r^3)\,\mathbf{r}$. The field $\mathbf{E}$ is potential at all points, except at the origin; here $\mathbf{E}$ can be written in the form $\mathbf{E} = -\operatorname{grad}(e/r)$. The quantity e/r is called the *potential of an electrostatic field.*

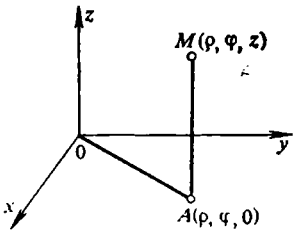


Fig. 29

Example 4°. Check to see that the field $\mathbf{F} = 2\rho z \sin\varphi\,\mathbf{e}_\rho + \rho^2 z \cos\varphi\,\mathbf{e}_\varphi + \rho^2 \sin\varphi\,\mathbf{e}_z$ specified in cylindrical coordinates is potential, and find its potential.

The field $\mathbf{F}$ is defined in the whole space. Let us check to see that $\operatorname{curl}\mathbf{F} = 0$:

$$\operatorname{curl}\mathbf{F} = \frac{1}{\rho}\begin{vmatrix} \mathbf{e}_\rho & \rho\mathbf{e}_\varphi & \mathbf{e}_z \\ \dfrac{\partial}{\partial\rho} & \dfrac{\partial}{\partial\varphi} & \dfrac{\partial}{\partial z} \\ 2\rho z\sin\varphi & \rho^2 z\cos\varphi & \rho^2\sin\varphi \end{vmatrix}$$

$$= \rho(\cos\varphi - \cos\varphi)\,\mathbf{e}_\rho - 2\rho(\sin\varphi - \sin\varphi)\,\mathbf{e}_\varphi + 2(z\cos\varphi - z\cos\varphi)\,\mathbf{e}_z = 0.$$

By virtue of Theorem 4, the field $\mathbf{F}$ is potential in the entire space. To compute the potential f by formula (7), we choose the origin for M_0 and take the broken line OAM for the path of integration (Fig. 29); we then have

$$f(M) = \int_{OA} \mathbf{F}\cdot d\mathbf{r} + \int_{AM} \mathbf{F}\cdot d\mathbf{r}.$$

Since on the path segment OA we have $\varphi = \text{const}$, $z = 0$, and $d\mathbf{r} = d\rho\,\mathbf{e}_\rho$, and on the path segment AM we have $\rho = \rho_{\text{fixed}}$, $\varphi = \varphi_{\text{fixed}}$, and $d\mathbf{r} = dz\,\mathbf{e}_z$, we obtain the following representation for the potential f:

$$f(M) = \int_0^\rho 0\,d\rho + \int_0^z \rho^2\sin\varphi\,dz = z\rho^2\sin\varphi + c.$$

Sec. 2.2. SOLENOIDAL VECTOR FIELD

1. Notion and Properties of a Solenoidal Field. A vector field $\mathbf{F}(M)$ is said to be *solenoidal* in a domain Ω if at each point of the field its divergence is equal to zero:

$$\operatorname{div} \mathbf{F} = 0. \tag{1}$$

The equality of the divergence to zero means that the solenoidal field is free of sources.

Let us consider the simplest properties of a solenoidal field.

A. *From the Ostrogradsky-Gauss formula it follows that if a solenoidal field is given in a simply connected domain, then the flux of the vector across any closed surface belonging to this domain is equal to zero:*

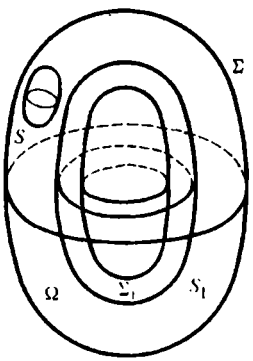


Fig. 30

$$\Pi = \oiint_{\Sigma} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma = \iiint_{\Omega} \operatorname{div} \mathbf{F} \, dv = 0. \tag{2}$$

For a solenoidal field given in a multiply connected domain statement (2) is, generally speaking, incorrect. This means that in a multiply connected domain there may exist closed surfaces the flux of the vector across which is different from zero.

Let us consider, for instance, the doubly connected domain Ω bounded by an outer Σ and an inner Σ_1 surface (Fig. 30). Let us take an arbitrary closed surface $S \subset \Omega$ not enveloping the surface Σ_1; the flux across this surface S is equal to zero, that is,

$$\Pi = \oiint_{S} \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma = 0.$$

And if the surface S_1 envelops the inner surface Σ_1, then the flux across it is, generally speaking, different from zero.

B. *Let a solenoidal field* F *be given in a simply connected domain. Then the flux of the vector* F *across any surface* Σ *spanned on a given contour* Γ *is independent of the kind of this surface and depends only on the contour* Γ.

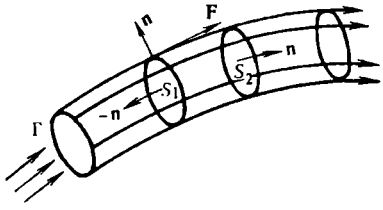


Fig. 31

Indeed, let us span two surfaces Σ_1 and Σ_2 on the contour Γ and apply to the domain Ω enclosed between these surfaces the Ostrogradsky-Gauss formula:

$$\iiint_{\Omega} \operatorname{div} \mathbf{F}\, dv = \iint_{\Sigma_1} \mathbf{F} \cdot \mathbf{n}^0\, d\sigma - \iint_{\Sigma_2} \mathbf{F} \cdot \mathbf{n}^0\, d\sigma.$$

The field is solenoidal, therefore $\operatorname{div} \mathbf{F} = 0$, and, hence, also $\iiint_{\Omega} \operatorname{div} \mathbf{F}\, dv = 0$. Consequently, the fluxes across the surfaces are also equal:

$$\iint_{\Sigma_1} \mathbf{F} \cdot \mathbf{n}^0\, d\sigma = \iint_{\Sigma_2} \mathbf{F} \cdot \mathbf{n}^0\, d\sigma,$$

that is, the flux is independent of the kind of the surfaces spanned on the contour Γ.

C. This property of a solenoidal field refers to the notion of the vector tube. Let us take a closed contour Γ in the field F and draw vector lines through its points (as in Fig. 31). The surface thus formed is called the *vector tube*. Any other vector line not passing through the points belonging to the contour Γ either lies wholly in the vector tube or is found outside of it. In the case of the velocity field of a stationary flux of a fluid, the vector tube is that part of space which is

filled by some fixed volume of the fluid, as the latter displaces.

The *intensity of the vector tube* is defined as the flux of the field across the cross-section of this tube. For solenoidal fields there takes place the so-called law of preserving the intensity of the vector tube.

If a solenoidal field F *is defined in a simply connected domain* Ω, *then the intensity of the vector tube is constant along the whole tube.*

Let us choose two arbitrary cross-sections of the tube: S_1 and S_2 (see Fig. 31). We then find the flux across the closed surface consisting of S_1^-, S_2^+, and the portion of the surface Σ of the vector tube enclosed between S_1 and S_2. By Property A, the flux of the vector F across this surface is equal to zero:

$$\oiint_{\Sigma+S_1^-+S_2^+} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma = \iint_{\Sigma} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma + \iint_{S_1^-} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma + \iint_{S_2^+} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma = 0.$$

The lateral surface Σ is formed by the vector lines of the field F, therefore the normal to the surface Σ will be a normal to the vector F as well, and then the flux across the lateral surface is equal to zero, that is,

$$\iint_{\Sigma} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma = 0.$$

The directions of the normal to S_1^- and S_2^+ are opposite. Shifting the integral over S_1^- leftwards and reversing the direction n in it (then the fluxes across S_1^+ and S_2^+ will be computed in one and the same direction), we obtain the equality

$$\iint_{S_2} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma = \iint_{S_1} \mathbf{F}\cdot\mathbf{n}^0\,d\sigma,$$

which means that the flux of the vector F across any section of the vector tube has one and the same value.

If a solenoidal field **F** is regarded as the velocity field of an incompressible fluid without sources and sinks, then Property C means that the quantity of fluid flowing per unit time across the section of the vector tube is one and the same for all sections of this tube.

D. *In a solenoidal field, the vector lines can neither begin nor terminate inside the field; they are either closed or have end points on the boundary of the field or have infinite branches (if the field is infinite).*

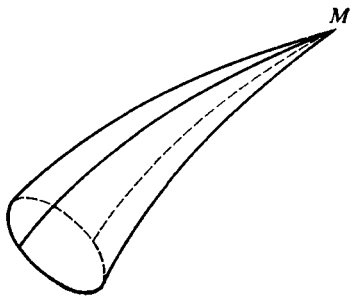


Fig. 32

Indeed, let the tube end at the point M (Fig. 32). By Property C, the intensity of the tube is constant everywhere, although the cross-section at the point M is equal to zero. Therefore at the point M the vector $\mathbf{F}(M)$ must attain an infinitely large value, which is impossible, since, by the supposition, the vector $\mathbf{F}(M)$ is continuous at every point. If we suppose that the tube terminates in the field with a finite section S, then at the points of this section the field will also be discontinuous, but this is impossible.

2. A Field of Sources and Sinks. In the preceding section (see Example 3°), it was shown that the vector field given by the expression

$$\mathbf{F} = \frac{q}{r^3}\mathbf{r} = \frac{q}{r^2}\mathbf{e}_r, \quad q = \text{const}, \tag{3}$$

is potential at all points, except for the origin. The vector field (3) is called the *field of a point source* ($q > 0$) or *of a sink* ($q < 0$), and the quantity q the *source intensity*.

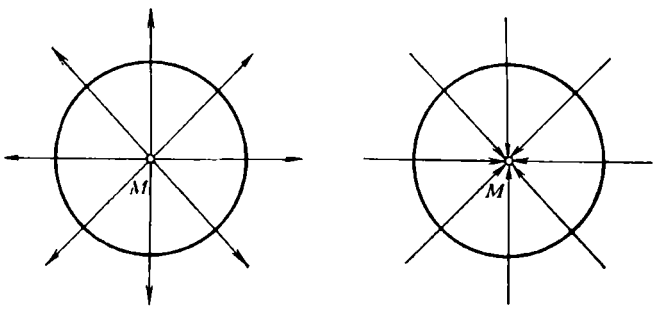
Let us characterize this field. The *vector lines of the field* are straight lines passing through a source or sink, the magnitude of the vector changing inversely proportional to the square of the distance of the point to a source or sink. The general character of the vector lines is shown in Fig. 33: (*a*) for a source and (*b*) for a sink.

Field (3) can be both a gravitational field formed by the mass $m = q/\gamma$ and an electrostatic field generated by the charge q.

Let us find the divergence of field (3). It is convenient to be computed in spherical coordinates. Taking advantage of formula (9) derived in Sec. 1.5, we get

$$\operatorname{div} \mathbf{F} = \frac{1}{r^2}\left(r^2 \frac{q}{r^2}\right)' = 0.$$

Thus, the field of a point source is solenoidal at all points except for the origin, where the vector field is not defined.



(*a*) (*b*)

Fig. 33

If Σ is an arbitrary surface not containing the origin inside itself, then, by Property A, the flux of vector (3) across this surface is equal to zero. And if Σ is a surface containing the origin inside itself, then the flux is, generally speaking, different from zero. Let us compute the flux of vector field (3) across the sphere S of radius R centered at the origin. On this sphere we have

$$|\mathbf{r}| = R, \quad \mathbf{F} = \frac{q}{R^3}\mathbf{r}, \quad \mathbf{n}^0 = \frac{\mathbf{r}}{R}, \quad \mathbf{F}\cdot\mathbf{n}^0 = \frac{q}{R^2}.$$

Consequently, the flux takes on the value

$$\iint\limits_S \mathbf{F}\cdot\mathbf{n}^0\, d\sigma = \frac{q}{R^2}\iint\limits_S d\sigma = 4\pi q.$$

Hence we see that the flux is independent of the radius of the sphere S, it is the same across any sphere enveloping the origin.

The field of one point source (sink) is readily generalized to obtain a field of k, point sources. Let k sources with intensities q_i, $i = 1, 2, \ldots, k$ be situated inside the sphere S. The flux of the vector field $\mathbf{F} = \sum_{i=1}^{k} \frac{q_i (\mathbf{r} - \mathbf{r}_i)}{|\mathbf{r} - \mathbf{r}_i|^3}$, where $\mathbf{r}_i$ are radius vectors of the sources, across the surface S is equal to the sum of the fluxes of point sources, i.e.

$$\Pi = \iint_S \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma = 4\pi \sum_{i=1}^{k} q_i. \tag{4}$$

The magnitude of the flux S remains the same if, instead of the sphere S, we take any other closed smooth oriented surface Σ enveloping the origin once.

Let us give a physical interpretation of the result obtained. Consider, in particular, an electrostatic field. Suppose the volume Ω bounded by the surface Σ is filled with charges with space density $\rho(M)$. Passing to the limit in equality (4), we can show that the flux of the electrostatic field $\mathbf{F}$ created by a certain distribution of the charges across the surface Σ is equal to the space charge in Ω multiplied by 4π, that is,

$$\oiint_\Sigma \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma = 4\pi \iiint_\Omega \rho \, dV.$$

Dividing both sides of the obtained equality by the voluume V and passing to the limit, we obtain

$$\lim_{\substack{\Sigma \to M \\ (V \to 0)}} \frac{\oiint_\Sigma \mathbf{F} \cdot \mathbf{n}^0 \, d\sigma}{v} = 4\pi \lim_{\substack{\Sigma \to M \\ (V \to 0)}} \frac{\iiint_\Omega \rho \, dv}{v} = 4\pi\rho(M)$$

or

$$\operatorname{div} \mathbf{F} = 4\pi\rho(M). \tag{5}$$

The last equality expresses Gauss' theorem in electrostatics: the divergence of the electrostatic field formed by some distribution of charges is equal to the space density $\rho(M)$ of the charges multiplied by 4π.

If a gravitational field is considered, then equality (5) means that the divergence of the gravitational field formed by a certain distribution of masses is equal to the space density $\rho(M)$ multiplied by 4π.

3. Vector Potential. Let us give a criterion by which a field $\mathbf{F}(M)$ will be solenoidal, using the notion of the vector potential.

The vector field $\mathbf{W}(M)$ is called the *vector potential of the field* $\mathbf{F}(M)$ if the field $\mathbf{F}(M)$ is representable in the form of the curl of the field $\mathbf{W}(M)$, that is, if

$$\mathbf{F}(M) = \operatorname{curl} \mathbf{W}(M).$$

The vector potential $\mathbf{W}(M)$ is determined with an accuracy to the gradient of an arbitrary scalar field $f(M)$. Indeed, if $\operatorname{curl} \mathbf{W}(M) = \mathbf{F}(M)$ and $f(M)$ is an arbitrary scalar field, then, using the relationship (6) from Sec. 1.6, we have

$$\operatorname{curl}(\mathbf{W}(M) + \operatorname{grad} f(M)) = \operatorname{curl} \mathbf{W}(M) + \operatorname{curl} \operatorname{grad} f(M) = \mathbf{F}(M).$$

Theorem 1. *In order for a continuously differentiable field* $\mathbf{F}(M)$ *to be solenoidal, it is necessary and sufficient that it have a vector potential* $\mathbf{W}(M)$.

The sufficiency of this condition follows directly from the relationship

$$\operatorname{div} \mathbf{F} = \operatorname{div}(\operatorname{curl} \mathbf{W}) = \nabla \cdot (\nabla \times \mathbf{W}) = (\nabla \times \nabla) \cdot \mathbf{W} = 0,$$

and the necessity will be the consequence of solvability of the system of differential equations

$$F_x = \frac{\partial W_z}{\partial y} - \frac{\partial W_y}{\partial z}, \quad F_y = \frac{\partial W_x}{\partial z} - \frac{\partial W_z}{\partial x}, \quad F_z = \frac{\partial W_y}{\partial x} - \frac{\partial W_x}{\partial y}$$

provided $\operatorname{div} \mathbf{F} = 0$. Without dwelling on the proof of solvability of this system in the general case, we are going to consider a particular example.

Example 1°. Show that if ω is a constant vector, then the vector field $\mathbf{F}(M) = r(\omega \times \mathbf{r})$ is solenoidal. Find one of the vector potentials of this field.

Using the properties of the operator ∇, we have

$$\operatorname{div} \mathbf{F} = \nabla \cdot (r(\omega \times \mathbf{r})) = (\nabla r) \cdot (\omega \times \mathbf{r}) + r\nabla \cdot (\omega \times \mathbf{r}). \quad (6)$$

Let us show that each term in representation (6) is equal to zero. We have $\nabla r = \operatorname{grad} r = \mathbf{r}/r$, and the first term equals zero as a scalar triple product with equal factors:

$$(\nabla r) \cdot (\omega \times \mathbf{r}) = \frac{1}{r} \mathbf{r} \cdot (\omega \times \mathbf{r}) = 0.$$

By Property G of the operator ∇ (see Sec. 1.6), we represent the second term in the form of the sum

$$r\nabla \cdot (\omega \times \mathbf{r}) = -r\omega \cdot (\nabla \times \mathbf{r}) + r\mathbf{r} \cdot (\nabla \times \omega),$$

each term in which is equal to zero, since $\nabla \times \mathbf{r} = \operatorname{curl} \mathbf{r} = 0$ and $\nabla \times \omega = 0$ (ω is a constant vector).

Thus, $\operatorname{div} \mathbf{F} = 0$ and the field $\mathbf{F}$ is solenoidal.

Let us now find one of the vector potentials of this field. If we point the z-axis along the direction of the vector ω, then the field $\mathbf{F}$ will be written in the following way:

$$\mathbf{F} = r \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & \omega \\ x & y & z \end{vmatrix} = r\omega(-y\mathbf{i} + x\mathbf{j}),$$

and then system (5) will be rewritten with respect to the vector $\mathbf{W}$ as follows:

$$\frac{\partial W_z}{\partial y} - \frac{\partial W_y}{\partial z} = -r\omega y, \quad \frac{\partial W_x}{\partial z} - \frac{\partial W_z}{\partial x} = r\omega x,$$

$$\frac{\partial W_y}{\partial x} - \frac{\partial W_x}{\partial y} = 0.$$

The vector potential $\mathbf{W}$ is determined with an accuracy to $\operatorname{grad} f$, therefore we may regard that $W_y = 0$, and then, from the third equation, we have the equality $\partial W_x/\partial y = 0$, i.e. $W_x = \psi(x, z)$. From the first equation of the system $\partial W_z/\partial y = -r\omega y$, we find

$$W_z = \omega \int y \sqrt{x^2 + y^2 + z^2}\, dy + f(x, z) = -\frac{\omega}{3} r^3 + f(x, z).$$

The functions W_x and W_z must satisfy the second equation of the system, therefore, on finding the partial derivatives:

$$\frac{\partial W_z}{\partial x} = -\omega x r + \frac{\partial f}{\partial x} \quad \text{and} \quad \frac{\partial W_x}{\partial z} = \frac{\partial \psi}{\partial z},$$

we see that the functions $\psi(x, z)$ and $f(x, z)$ must satisfy the condition $\partial\psi/\partial z - \partial f/\partial x = 0$. Setting, in particular, $\psi(x, z) = f(x, z) = 0$, we find one of the vector potentials:

$$\mathbf{W} = -\frac{\omega}{3} r^3 \mathbf{k} = -\frac{1}{3} r^3 \boldsymbol{\omega}.$$

Sec. 2.3. LAPLACE'S VECTOR FIELD

1. Differential Operations of the Second Order. The operations of finding gradient, curl, and divergence may be called the differential operations of the first order. Let us now consider the basic differential operations of the second order. The operations grad f and curl $\mathbf{F}$ are vectors, therefore we may apply to them the operations of finding divergence and curl, while to div $\mathbf{F}$ only one operation is applicable—that of finding gradient. Thus, we obtain five operations of the second order having the form

$$\begin{gathered} \operatorname{div} \operatorname{grad} f = \nabla \cdot (\nabla f), \quad \operatorname{curl} \operatorname{grad} f = \nabla \times (\nabla f), \\ \operatorname{div} \operatorname{curl} \mathbf{F} = \nabla \cdot (\nabla \times \mathbf{F}), \quad \operatorname{curl} \operatorname{curl} \mathbf{F} = \nabla \times (\nabla \times \mathbf{F}); \quad (1) \\ \operatorname{grad} \operatorname{div} \mathbf{F} = \nabla (\nabla \cdot \mathbf{F}). \end{gathered}$$

Two of these operations (the second and third) yield zero, since $\nabla \cdot (\nabla \times \mathbf{F}) = 0$ is a triple scalar product of the vectors whose two factors are equal, and $\nabla \times (\nabla f) = 0$ is a vector product with equal factors. Applying the equality $\mathbf{a} \times \mathbf{b} \times \mathbf{c} = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$ for a triple vector product to the vectors ∇ and $\mathbf{F}$, we get

$$\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - (\nabla \cdot \nabla)\mathbf{F}. \quad (2)$$

From this expression we conclude that the last two operations (1) are related by (2).

The scalar square of the nabla vector, that is, the expression

$$\nabla \cdot \nabla = \nabla^2 = \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}\right)^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

is called the *Laplacian operator* (or *Laplacian*) and is denoted by Δ. Like the operator ∇, this operator is also widely used.

The symbolic Laplacian operator may be applied to both the scalar function $f(M)$ and the vector function $\mathbf{F}(M)$. By this application we shall understand the equalities

$$\Delta f = \nabla \cdot (\nabla f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \tag{3}$$

and

$$\Delta \mathbf{F} = \Delta F_x \mathbf{i} + \Delta F_y \mathbf{j} + \Delta F_z \mathbf{k}.$$

Of the differential operations of the second order, the most widely used is the first operation which is written, with the aid of the Laplacian operator, in the following way: div grad $f = \Delta f$, and is also called the Laplacian.

To find the expression of the Laplacian in an arbitrary curvilinear coordinate system, let us take advantage of the expressions grad f and div $\mathbf{F}$ in this system. In formula (7) derived in Sec. 1.5 we replace the vector $\mathbf{F}$ by grad f determined by equality (9) from Sec. 1.2 and thus obtain the expression for the Laplacian in the curvilinear orthogonal coordinate system:

$$\Delta f = \frac{1}{H_1 H_2 H_3} \left[\frac{\partial}{\partial q^1} \left(\frac{H_2 H_3}{H_1} \frac{\partial f}{\partial q^1} \right) + \frac{\partial}{\partial q^2} \left(\frac{H_3 H_1}{H_2} \frac{\partial f}{\partial q^2} \right) + \frac{\partial}{\partial q^3} \left(\frac{H_1 H_2}{H_3} \frac{\partial f}{\partial q^3} \right) \right]. \tag{4}$$

In particular, for cylindrical coordinates we shall have the expression

$$\Delta f = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial f}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 f}{\partial \varphi^2} + \frac{\partial^2 f}{\partial z^2}, \tag{5}$$

and for spherical coordinates the expression

$$\Delta f = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \varphi^2} + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial f}{\partial \theta} \right). \tag{6}$$

For a plane scalar field the most widely used are the Cartesian and polar coordinate systems. In the Cartesian co-

ordinate system the Laplacian is written in the form

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}, \tag{7}$$

and in the polar coordinate system in the form

$$\Delta f = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \frac{\partial f}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 f}{\partial \varphi^2}. \tag{8}$$

2. Green's Formulas. Consider the application of the Ostrogradsky-Gauss formula to a special kind of the vector field formed with the aid of two scalar functions f and ψ:

$$\mathbf{F} = f \operatorname{grad} \psi.$$

For the divergence of this field we have the expression

$$\operatorname{div} \mathbf{F} = \nabla \cdot (f \nabla \psi) = \nabla f \cdot \nabla \psi + f \nabla^2 \psi$$
$$= \operatorname{grad} f \cdot \operatorname{grad} \psi + f \Delta \psi,$$

and the scalar product is written in the form

$$\mathbf{F} \cdot \mathbf{n}^0 = f \operatorname{grad} \psi \cdot \mathbf{n}^0 = f \frac{d\psi}{dn}.$$

The Ostrogradsky-Gauss formula (7) (Sec. 1.4) then takes the form

$$\iint_{\Sigma} f \frac{d\psi}{dn} d\sigma = \iiint_{\Omega} (\operatorname{grad} f \cdot \operatorname{grad} \psi + f \Delta \psi)\, dv. \tag{9}$$

Formula (9) is usually called *Green's first formula.*

Interchanging the functions f and ψ, we obtain the equality

$$\iint_{\Sigma} \psi \frac{df}{dn} d\sigma = \iiint_{\Omega} (\operatorname{grad} f \cdot \operatorname{grad} \psi + \psi \Delta f)\, dv;$$

subtracting it from (9), we get *Green's second formula*:

$$\iint_{\Sigma} \left(f \frac{d\psi}{dn} - \psi \frac{df}{dn}\right) d\sigma = \iiint_{\Omega} (f \Delta \psi - \psi \Delta f)\, dv. \tag{10}$$

Putting in Green's first formula (9) $f = \psi$, we find *Green's third formula*:

$$\oiint_{\Sigma} f \frac{df}{dn} d\sigma = \iiint_{\Omega} ((\operatorname{grad} f)^2 + f \Delta f)\, dv. \qquad (11)$$

Green's formulas also hold for the case of two variables. Let τ be a tangent and n a normal at an arbitrary point of a smooth contour γ. We denote by α the angle between the x-axis and the tangent τ. Then the unit vector of the tangent will be represented in the form $\tau^0 = \cos\alpha\, i + \sin\alpha\, j$, and the unit vector of the normal in the form $n^0 = \sin\alpha\, i - \cos\alpha\, j$. Applying Green's formula (14) from Sec. 1.3 to the integral $\oint_{\gamma} F \cdot n^0\, dl$, we obtain the equalities

$$\oint_{\gamma} F \cdot n^0\, dl = \oint_{\gamma} -F_x\, dy + F_y\, dx = \iint_{G} \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y}\right) dx\, dy.$$

Setting here $F = f \operatorname{grad} \psi$, we find Green's first formula for the case of two variables:

$$\oint_{\gamma} f \frac{d\psi}{dn} dl = \iint_{G} (\operatorname{grad} f \cdot \operatorname{grad} \psi + f \Delta\psi)\, d\sigma. \qquad (12)$$

Green's second and third formulas are obtained in the same way as in the case of three variables and have, respectively, the form

$$\oint_{\gamma} \left(f \frac{d\psi}{dn} - \psi \frac{df}{dn}\right) dl = \iint_{G} (f \Delta\psi - \psi \Delta f)\, d\sigma, \qquad (13)$$

$$\oint_{\gamma} f \frac{df}{dn} dl = \iint_{G} ((\operatorname{grad} f)^2 + f \Delta f)\, d\sigma. \qquad (14)$$

Thus, Green's formulas for the cases of two and three variables have the same form.

3. Harmonic Functions. A vector field F is called *Laplacian* if it is potential and solenoidal at the same time. The potentiality of Laplace's field F implies the existence of a potential f such that $F = \operatorname{grad} f$. From the solenoidality of Laplace's field we have $\operatorname{div} F = 0$, therefore $\operatorname{div} \operatorname{grad} f = \Delta f = 0$.

As distinct from an arbitrary vector field **F** defined by three scalar functions, a Laplace's vector field is defined by one scalar function—the potential f, which is the solution of the equation

$$\Delta f = 0 \tag{15}$$

called *Laplace's equation.*

The function f having continuous partial derivatives up to the second order inclusively and satisfying Laplace's equation (15) is called *harmonic.*

The harmonic function *in space* is exemplified by the function $f = 1/r$ defined everywhere except at $r = 0$. Indeed, taking advantage of the expression for Laplacian (6) in spherical coordinates, we get the equalities

$$\Delta \frac{1}{r} = \frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d(1/r)}{dr} \right) = \frac{1}{r^2} \frac{d(-1)}{dr} = 0.$$

The function $f = \ln(1/\rho)$ is a harmonic function *in the plane.* Indeed, applying formula (8), we obtain the equalities

$$\Delta \ln \frac{1}{\rho} = \frac{1}{\rho} \frac{d}{d\rho} \left(\rho \frac{d}{d\rho} \left(\ln \frac{1}{\rho} \right) \right) = \frac{1}{\rho} \frac{d(-1)}{d\rho} = 0.$$

The gravitational field $\mathbf{F} = -\frac{\gamma m}{r^3} \mathbf{r}$ is an example of a Laplace's vector field. It was shown in the previous sections that this field is potential and the potential f is equal to $\gamma m/r$. But the function $1/r$ is harmonic, therefore the gravitational field is Laplacian everywhere except at the point $r = 0$.

For Laplace's field **F** given in a simply connected domain, as it follows from the definition and Theorem 4 proved in Sec. 2.1, the equalities

$$\operatorname{div} \mathbf{F} = 0, \quad \operatorname{curl} \mathbf{F} = 0$$

are fulfilled simultaneously.

4. Integral Representation of a Function. Let us derive a formula expressing the value of the function $f(P)$ at an arbitrary point $P \in \Omega$ with the aid of a triple and surface integrals called the *integral representation of a function.*

Theorem 1. *If the function $f(M)$ and the partial derivatives up to the second order inclusively are continuous in a finite three-dimensional closed domain $\bar{\Omega}$ bounded by a piecewise*

smooth surface Σ, *then*

$$f(P)=\frac{1}{4\pi}\left(\oint\!\!\!\oint_{\Sigma}\frac{1}{r}\frac{\mathrm{d}f}{\mathrm{d}n}\,d\sigma\right.$$

$$\left.-\oint\!\!\!\oint_{\Sigma} f\frac{\mathrm{d}}{\mathrm{d}n}\frac{1}{r}\,\mathrm{d}\sigma-\iiint_{\Omega}\frac{1}{r}\,\Delta f\,\mathrm{d}v\right), \qquad (16)$$

where r *is the distance between a fixed point* P *and an arbitrary point* $M\in\bar{\Omega}$.

The function $1/r=1/r_{PM}$ is harmonic everywhere in the domain Ω except at the point P. Let us enclose the point P in a ball G of radius ε centred at this point. The surface of the ball will be denoted by Σ_ε. The function $1/r_{PM}$ will be harmonic in the domain $\Omega-G$. Applying Green's second formula (10) to the functions f and $\psi=1/r$ in the domain $\Omega-G$, we obtain the equality

$$\iint_{\Sigma+\Sigma_\varepsilon^-}\left(f\frac{\mathrm{d}}{\mathrm{d}n}\left(\frac{1}{r}\right)-\frac{1}{r}\frac{\mathrm{d}f}{\mathrm{d}n}\right)\mathrm{d}\sigma=-\iiint_{\Omega-G}\frac{1}{r}\,\Delta f\,\mathrm{d}v. \qquad (17)$$

The normal to the inner side of the sphere Σ_ε^- is directed along the radius r of the sphere towards its centre, therefore the normal derivative has the form

$$\frac{\mathrm{d}(1/r)}{\mathrm{d}n}=-\frac{\mathrm{d}(1/r)}{\mathrm{d}r}=\frac{1}{r^2}.$$

Applying the mean-value theorem to the surface integrals

$$I_1=\iint_{\Sigma_\varepsilon^-} f\frac{\mathrm{d}(1/r)}{\mathrm{d}n}\,\mathrm{d}\sigma \quad\text{and}\quad I_2=\iint_{\Sigma_\varepsilon^-}\frac{1}{r}\frac{\mathrm{d}f}{\mathrm{d}n}\,\mathrm{d}\sigma$$

and taking into consideration that on the surface of the sphere $r=\varepsilon$, we shall have the following representation of the integrals:

$$I_1=\iint_{\Sigma_\varepsilon} f\frac{1}{\varepsilon^2}\,\mathrm{d}\sigma=\frac{1}{\varepsilon^2}\iint_{\Sigma_\varepsilon} f\,\mathrm{d}\sigma=\frac{f(M_1)}{\varepsilon^2}\,4\pi\varepsilon^2=4\pi f(M_1),$$

$$I_2=-\frac{1}{\varepsilon}\iint_{\Sigma_\varepsilon}\frac{\mathrm{d}f}{\mathrm{d}n}\,\mathrm{d}\sigma=-\frac{1}{\varepsilon}\left(\frac{\mathrm{d}f}{\mathrm{d}n}\right)_{M_2}4\pi\varepsilon^2=-4\pi\varepsilon\left(\frac{\mathrm{d}f}{\mathrm{d}n}\right)_{M_2},$$

where M_1 and M_2 are points of the sphere Σ_ε. Substituting I_1 and I_2 into equality (17), we rewrite it in the form

$$f(M_1) = \frac{1}{4\pi}\left(\iint\limits_{\Sigma} \frac{1}{r}\frac{df}{dn}\,d\sigma - \iint\limits_{\Sigma} f\frac{d(1/r)}{dn}\,d\sigma - \iiint\limits_{\Omega-G}\frac{1}{r}\Delta f\,dv + 4\pi\varepsilon\left(\frac{df}{dn}\right)_{M_2}\right). \quad (18)$$

Let us pass to the limit in the last equality as $\varepsilon \to 0$. Here the surface integrals are independent of ε, the expression $4\pi\varepsilon\left(\frac{df}{dn}\right)_{M_2} \to 0$ as $\varepsilon \to 0$, and it is possible to show that the improper integral $\iiint\limits_{\Omega}\frac{1}{r}\Delta f\,dv$ exists and is the limit of the integral $\iiint\limits_{\Omega-G}\frac{1}{r}\Delta f\,dv$. Thus, passing to the limit in equality (18), we obtain formula (16) which yields the integral representation of the function $f(P)$.

In the particular case, when f is a harmonic function, i.e. $\Delta f = 0$, the triple integral in formula (16) disappears and we obtain for the harmonic function the integral representation in the form of a linear combination of surface integrals:

$$f(P) = \frac{1}{4\pi}\left(\iint\limits_{\Sigma}\frac{1}{r}\frac{df}{dn}\,d\sigma - \iint\limits_{\Sigma} f\frac{d(1/r)}{dn}\,d\sigma\right). \quad (19)$$

A formula analogous to (16) can also be obtained for a plane domain G. This formula is given in the following theorem.

Theorem 2. *If the function $f(M)$ and the partial derivatives up to the second order inclusively are continuous in a plane closed domain $\bar{G}$ bounded by the contour γ, then the value of f at the interior point P is determined by the formula*

$$f(P) = \frac{1}{2\pi}\left(\oint\limits_{\gamma}\ln\frac{1}{r}\frac{df}{dn}\,dl - \oint\limits_{\gamma} f\frac{d\ln(1/r)}{dn}\,dl - \iint\limits_{G}\ln\frac{1}{r}\Delta f\,d\sigma\right), \quad (20)$$

where r is the distance between the points P and $M \in \bar{G}$.

This theorem is proved similarly to Theorem 1 by using Green's second formula for the case of two variables (13) in which the function $\psi = \ln(1/r)$ is harmonic everywhere except at the point $r = 0$.

In a particular case, when f is a harmonic function, formula (20) takes the form

$$f(P) = \frac{1}{2\pi}\left(\oint_{\gamma} \ln\frac{1}{r}\frac{df}{dn}\,dl - \oint_{\gamma} f\frac{d\ln(1/r)}{dn}\,dl\right). \qquad (21)$$

5. Properties of Harmonic Functions. Consider the function f, harmonic in a simply connected three-dimensional domain Ω. Let us enumerate its properties.

A. *The integral of the normal derivative of a harmonic function over any closed surface* Σ *lying in the domain* Ω *is equal to zero*:

$$\iint_{\Sigma} \frac{df}{dn}\,d\sigma = 0. \qquad (22)$$

Equality (22) is obtained from Green's second formula (10) in which the function $\psi = 1$.

B. **Theorem 3 (Mean-value Theorem).** The value of a harmonic function at the centre P of the spherical domain G_R is equal to its arithmetic mean value on the sphere Σ_R (R—radius of the sphere):

$$f(P) = \frac{1}{4\pi R^2}\iint_{\Sigma_R} f(M)\,d\sigma. \qquad (23)$$

On the sphere Σ_R with centre at the point P we have: $r = R$, $d(1/r)/dn = -1/R^2$. Applying first formula (19) and then taking advantage of Property A of harmonic functions, we obtain the equalities which prove formula (23):

$$f(P) = \frac{1}{4\pi}\left(\frac{1}{R}\iint_{\Sigma_R}\frac{df}{dn}\,d\sigma + \frac{1}{R^2}\iint_{R^2} f\,d\sigma\right) = \frac{1}{4\pi R^2}\iint_{\Sigma_R} f\,d\sigma.$$

C. **Theorem 4.** *If the function* f, *harmonic in the domain* Ω, *is continuous in the closed domain* $\bar{\Omega}$ *and is not constant, then it reaches its greatest and least values on the boundary of the domain.*

The function f is continuous in the closed bounded domain $\bar{\Omega}$, therefore it necessarily reaches its greatest and least values in this domain. Let us carry out the proof for the greatest value. Suppose the function f reaches its greatest value at the interior point M_0, i.e.

$$\max_{M \in \bar{\Omega}} f(M) = f(M_0).$$

The point M_0 may be not unique. But since the function $f(M)$ is not constant, for $M_0 \in \Omega$ we take that point in the neighbourhood of which there is a point M such that the inequality $f(M) < f(M_0)$ is fulfilled. Let us construct a sphere Σ_1 of radius $M_0M = R$ with centre at M_0 (Fig. 34). The continuity of the function $f(M)$ implies that the inequality $f(M) < f(M_0)$ will be fulfilled on some part of the sphere $\Sigma_1 \subset \Sigma$. By the mean-value theorem, for the sphere Σ we have

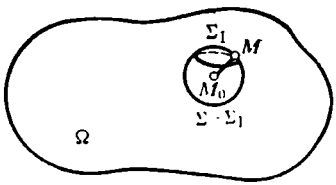


Fig. 34

$$f(M_0) = \frac{1}{4\pi R^2} \iint_{\Sigma} f(M)\, d\sigma$$

$$= \frac{1}{4\pi R^2} \left(\iint_{\Sigma_1} f\, d\sigma + \iint_{\Sigma - \Sigma_1} f\, d\sigma \right). \quad (24)$$

Fulfilled on Σ_1 is the strict inequality $f(M) < f(M_0)$, and on $\Sigma - \Sigma_1$ the slack inequality $f(M) \leqslant f(M_0)$, therefore from equality (24) we obtain the relationship

$$f(M_0) < \frac{f(M_0)}{4\pi R^2} \left(\iint_{\Sigma_1} d\sigma + \iint_{\Sigma - \Sigma_1} d\sigma \right) = \frac{f(M_0)}{4\pi R^2} \iint_{\Sigma} d\sigma = f(M_0).$$

Thus, we have come to a contradiction: $f(M_0) < f(M_0)$, and this just proves the theorem for the case of the greatest value. The statement for the least value is proved in a similar way.

Corollary. *If a harmonic in* Ω *and continuous in* $\bar{\Omega}$ *function* $f(M)$ *reaches the greatest (least) value inside the domain* Ω, *then* $f(M) \equiv \text{const}$.

If we suppose that the function $f(M) \neq \text{const}$ and reaches the greatest value at an interior point M_0, then we shall come to a contradiction with Theorem 4.

D. **Theorem 5.** *The function* $f(M)$, *harmonic in the bounded domain* Ω, *continuous in the closed domain* $\bar{\Omega}$, *and attaining constant values* c *on the boundary* Σ *of this domain, is identically equal to* c *in the domain* $\bar{\Omega}$.

The proof of this assertion follows directly from Property C. If $f \equiv c$ on the boundary Σ, then

$$\max_{M \in \bar{\Omega}} f(M) = \min_{M \in \bar{\Omega}} f(M) = c.$$

Hence, $f(M) \equiv c = \text{const}$ everywhere in the domain $\bar{G}$.

The above properties are also fulfilled for a function $f(M)$, harmonic in a plane domain G. Of course, it is natural that Properties A and B look differently, namely: formulas (22) and (23) are replaced by the respective equalities:

$$\oint_{\gamma} \frac{df}{dn}\, dl = 0 \tag{25}$$

and

$$f(P) = \frac{1}{2\pi R} \oint_{\gamma} f(M)\, dl. \tag{26}$$

Sec. 2.4. DIRICHLET PROBLEM AND NEUMANN PROBLEM

1. Statement of Boundary-value Problems, Their Uniqueness. In Chapter 1, we mainly studied a given scalar $f(M)$ or vector $\mathbf{F}(M)$ field. Introduced and studied in these fields were various differential operations. Sections 2.1 and 2.2 were dedicated to restoring a scalar or a vector field from their differential properties. Let us now solve the problem of finding a scalar field from its Laplacian:

$$\Delta f = g(M). \tag{1}$$

Equation (1) is called *Poisson's equation*, its particular case (for $g(M) = 0$) being termed Laplace's equation:

$$\Delta f = 0. \tag{2}$$

Poisson's equation (1) is a partial differential equation, and it is satisfied by infinitely many functions differing from one another by a harmonic function. For instance, the functions $f_1 = x^2/2$, $f_2 = y^2/2$, $f_3 = x^2/2 + 5xy$ and many other functions are solutions of the equation

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 1.$$

The problem consisting in finding a solution of Poisson's equation (1) not satisfying some additional conditions is of little interest. The role of additional conditions may be played by so-called boundary conditions.

Depending on boundary conditions various boundary-value problems are considered. Let there be given a domain Ω bounded by a surface Σ. If a solution is sought for inside the domain Ω, then one speaks of an inner boundary-value problem and if a solution is looked for outside of the domain Ω, then an outer boundary-value problem is meant. We will consider inner boundary-value problems and call them simply *boundary-value problems*. A boundary-value problem is set in the following way.

It is required to find such a function $f(M)$ which must: (a) be continuous in the closed domain $\bar{\Omega} = \Omega + \Sigma$; (b) have continuous partial derivatives up to the second order inclusively in the open domain Ω; (c) satisfy equation (1); (d) satisfy boundary conditions on Σ.

The Dirichlet and Neumann problems are the most important boundary-value problems.

The *Dirichlet problem* is formulated in the following way: find the solution of Poisson's equation (1) in the domain Ω attaining on the boundary the given values

$$f|_{\Sigma} = \alpha(N). \tag{3}$$

And if it is required to find the solutions of equation (1) in the domain Ω when the boundary values of the normal

derivative are given

$$\left.\frac{df}{dn}\right|_{\Sigma} = \beta(N), \tag{4}$$

then in this case we have the *Neumann problem.*

Without touching the questions of existence of the solutions of Dirichlet and Neumann problems, let us establish the uniqueness of these solutions. We shall assume that the function $g(M)$ is continuous in the region Ω, and the functions $\alpha(N)$ and $\beta(N)$ are continuous on the surface Σ, and that there exist the solutions $f(M)$ of the set boundary-value poblems.

Theorem 1. *The solution $f(M)$ of the Dirichlet problem, if it exists, is unique and continuously depends on the given boundary values* (3).

Suppose there exist two solutions of the Dirichlet problem: f_1 and f_2. Then the function $f = f_2 - f_1$ satisfies Laplace's equation (2) $\Delta f = \Delta f_2 - \Delta f_1 = g - g = 0$. Therefore the function f is harmonic in the domain Ω and continuous in the closed domain $\bar{\Omega}$. On the boundary Σ of the domain Ω the function f satisfies the zero boundary condition

$$f|_{\Sigma} = f_2|_{\Sigma} - f_1|_{\Sigma} = \alpha - \alpha = 0.$$

By virtue of Property D of harmonic functions given in the preceding section, it follows that $f \equiv 0$ in Ω, i.e. $f_1 \equiv f_2$. Thus the uniqueness has been established.

Let us pass over to the proof of the continuous dependence of the solution of the Dirichlet problem on the boundary conditions. It is necessary to show that to a small change of the boundary conditions (3) there corresponds a small change of the solution itself. Let f_1 be the solution of equation (1) satisfying the boundary condition $f_1|_{\Sigma} = \alpha_1(N)$, and f_2 the solution of equation (1) satisfying the condition $f_2|_{\Sigma} = \alpha_2(N)$, the boundary conditions $\alpha_1(N)$ and $\alpha_2(N)$ differing but little, that is,

$$\max_{N \in \Sigma} |\alpha_1(N) - \alpha_2(N)| < \varepsilon.$$

The function $f = f_1 - f_2$ satisfies Laplace's equation $\Delta f = 0$ and the boundary condition $f|_{\Sigma} = \alpha_1(N) - \alpha_2(N)$. But

for all points $M \in \Omega$ there holds the inequality

$$\min_{N \in \Sigma} f(N) \leqslant f(M) \leqslant \max_{N \in \Sigma} f(N),$$

from which there follows the inequality for the modulus of the function $f(M)$

$$|f(M)| \leqslant \max_{N \in \Sigma} |f(N)| = \max_{N \in \Sigma} |\alpha_1(N) - \alpha_2(N)| < \varepsilon.$$

Thus, $\forall M \in \Omega$ the solutions $f_1(M)$ and $f_2(M)$ differ by a quantity not exceeding ε by modulus, that is, $|f_1(M) - f_2(M)| < \varepsilon$.

The proved property of continuous dependence of the solution of the Dirichlet problem on boundary conditions is very important when solving physical problems, since small changes in the boundary conditions can cause only small changes in the solution of a problem.

Theorem 2. *The solution $f(M)$ of the Neumann problem, if it exists, is unique with an accuracy to an arbitrary constant.*

Let in the domain Ω there exist two solutions of the Neumann problem: f_1 and f_2. Then the function $f = f_1 - f_2$ satisfies Laplace's equation $\Delta f = 0$ and zero boundary condition $\left.\frac{df}{dn}\right|_{\Sigma} = 0$. Applying Green's third formula, we obtain the equality

$$\int\int\int_{\Omega} (\operatorname{grad} f)^2 \, dv = \oint\oint_{\Sigma} f \frac{df}{dn} \, d\sigma = 0.$$

Here,

$$(\operatorname{grad} f)^2 \geqslant 0, \quad \text{and} \quad \int\int\int_{\Omega} (\operatorname{grad} f)^2 \, dv = 0,$$

which is possible only in the case when $\operatorname{grad} f = 0$. By Property D of gradient (see Sec. 1.2) we conclude that $f = c = \text{const}$, that is, $f_2 = f_1 + c$.

2. Solving the Dirichlet Problem with the Aid of Green's Function. Formula (16) given in the preceding section is inconvenient, since it requires a simultaneous knowledge of the function $f(N)$ and its normal derivative df/dn on the boundary Σ. Consequently, in such a form, it cannot be used for solving the Dirichlet or the Neumann problem. Let

us transform the mentioned formula (16). We represent Green s second formula (10) from Sec. 2.3 (regarding the function $\psi(P, M) = \Phi(P, M)$ as harmonic) in the form

$$0 = \frac{1}{4\pi}\left(-\iint\limits_{\Sigma} f \frac{d\Phi}{dn}\, d\sigma + \iint\limits_{\Sigma} \Phi \frac{df}{dn}\, d\sigma - \iiint\limits_{\Omega} \Phi \Delta f\, dv\right)$$

and add this to formula 16 from Sec. 2.3:

$$f(P) = \frac{1}{4\pi}\left(\iint\limits_{\Sigma}\left(\frac{1}{r}+\Phi\right)\frac{df}{dn}\, d\sigma - \iint\limits_{\Sigma} f \frac{d}{dn}\left(\frac{1}{r}+\Phi\right) d\sigma - \iiint\limits_{\Omega}\left(\frac{1}{r}+\Phi\right)\Delta f\, dv\right).$$

If we introduce the notation

$$G(P, M) = \frac{1}{r}+\Phi = \frac{1}{r_{MP}}+\Phi(P, M), \tag{5}$$

then the preceding formula takes the form

$$f(P) = \frac{1}{4\pi}\left(\iint\limits_{\Sigma} G \frac{df}{dn}\, d\sigma - \iint\limits_{\Sigma} f \frac{dG}{dn} - \iiint\limits_{\Omega} G \Delta f\, dv\right). \tag{6}$$

Formula (6) holds true for any harmonic function $\Phi(P, M)$. A harmonic function is uniquely determined by its boundary values. Let us choose a harmonic function $\Phi(P, M)$ so that on the boundary Σ the equality

$$\Phi(P, M)\,|_{\Sigma} = \Phi(P, N) = -\frac{1}{r_{NP}}$$

is fulfilled, and the introduced function (5) vanishes

$$G(P, M)\,|_{\Sigma} = G(P, N) = 0. \tag{7}$$

Here, P is a fixed point of the domain Ω, and N is an arbitrary point of the boundary Σ. The function $G(P, M)$ defined by equality (5) and vanishing on the boundary Σ is called *Green's function of the Dirichlet problem.*

The solution of the Dirichlet problem applying Green's function consists of two stages:

(1) construction of Green's function (5) which is reduced to finding the harmonic function $\Phi(P, M)$ from its boundary conditions (7);

(2) direct computation of the solution of the Dirichlet problem by the formula

$$f(P) = -\frac{1}{4\pi} \oint\limits_{\Sigma} \alpha(N) \frac{dG}{dn} d\sigma - \frac{1}{4\pi} \int\limits_{\Omega} \int\limits_{\Omega} \int\limits_{\Omega} Gg \, dv. \quad (8)$$

Note that Green's function is determined only by the type of the surface Σ and is related neither to the boundary conditions (3) of the Dirichlet problem nor to the value of the Laplacian Δf in the domain Ω. Therefore if for a given domain Ω Green's function $G(P, M)$ is constructed, then it gives the solution of Poisson's equation (1) to the whole class of problems with arbitrary boundary values (3) and with an arbitrary right-hand side g. In particular, the solution of the Dirichlet problem for Laplace's equation (2) is given by the formula

$$f(P) = -\frac{1}{4\pi} \oint\limits_{\Sigma} \alpha(N) \frac{dG(P, N)}{dn} d\sigma(N). \quad (9)$$

Green's function (5) is also known as the *function of a point source*. Such a name is related to a physical interpretation of Green's function. Thus, in the case of an electrostatic field, the first term $1/r$ is the potential of a point charge, and the second term $\Phi(P, M)$ means the potential of the charge field induced on the conducting surface Σ. Thus, the construction of a source function is reduced to the determination of the induced field.

In the case of a plane domain D, Green's function has the form

$$G(P, M) = \ln \frac{1}{r_{PM}} + \Phi(P, M), \quad (10)$$

where $\Phi(P, M)$ is a harmonic function. The solution of the Dirichlet problem is given by the formula

$$f(P) = -\frac{1}{2\pi} \oint\limits_{\gamma} \alpha(N) \frac{dG}{dn} dl - \frac{1}{2\pi} \int\limits_{D} \int g(M) G(P, M) d\sigma \quad (11)$$

which is obtained from formula (20) given in the preceding section. The harmonic function $\Phi(P, M)$ in equality (10) is determined from the boundary condition

$$\Phi|_{\gamma} = \Phi(P, N) = -\ln \frac{1}{r_{PM}}\Big|_{\gamma} = -\ln \frac{1}{r_{PN}},$$

where P is the point under consideration, and N is a running point of the boundary γ. Formula (11) is obtained in the same way as formula (8). The reader is invited to derive this formula.

3. Solving the Neumann Problem with the Aid of Green's Function. The solution of the Neumann problem for Laplace's equation exists not for all boundary conditions (4) but only for those which satisfy the equality

$$\iint_{\Sigma} \beta \, d\sigma = \iint_{\Sigma} \frac{df}{dn} \, d\sigma = 0 \tag{12}$$

(see Property A of harmonic functions, Sec. 2.3).

To construct Green's function from formula (5), it is necessary to determine the harmonic function $\Phi(P, M)$. Note that there is no harmonic function $\Phi(P, M)$ satisfying the condition

$$\frac{d\Phi}{dn}\Big|_{\Sigma} = -\frac{d(1/r)}{dn}\Big|_{\Sigma}. \tag{13}$$

Indeed, setting in formula (16) from Sec. 2.3 $f(M) \equiv 1$, we get the relationship

$$\iint_{\Sigma} \frac{d(1/r)}{dn} \, d\sigma = -4\pi,$$

that is, for the function $d(1/r)/dn$ condition (12) is not fulfilled, and this just means that the harmonic function $\Phi(P, M)$, for which condition (13) is fulfilled, is inexistent. Therefore, instead of condition (13), we take the condition

$$\frac{d\Phi}{dn}\Big|_{\Sigma} = -\frac{d(1/r)}{dn}\Big|_{\Sigma} = -\frac{4\pi}{S(\Sigma)}, \tag{14}$$

where $S(\Sigma)$ is the area of the surface Σ. In this case condition (12) is fulfilled and it is possible to construct the harmonic function $\Phi(P, M)$ satisfying condition (14).

Green's function $G(P, M)$ is constructed by formula (5) and on the boundary Σ satisfies the condition

$$\frac{dG}{dn}\Big|_{\Sigma} = -\frac{4\pi}{S(\Sigma)}.$$

From equality (6) it follows that the solution of the Neumann problem is given by the formula

$$f(P) = \frac{1}{4\pi}\left(\iint\limits_{\Sigma} G(P, M)\beta(N)\,d\sigma - \iiint\limits_{\Omega} gG\,dv\right) - \frac{1}{S(\Sigma)}\iint\limits_{\Sigma} f\,d\sigma.$$

The integral $\frac{1}{S(\Sigma)}\iint\limits_{\Sigma} f\,d\sigma$ is the mean value of the unknown function; although the value itself is unknown, nevertheless it is constant. By Theorem 2, the solution of the Neumann problem is determined with an accuracy to a constant, therefore the solution of the set problem is written in the form

$$f(P) = \frac{1}{4\pi_4}\left(\iint\limits_{\Sigma} G(P, N)\beta(N)\,d\sigma - \iiint\limits_{\Omega} g(M)G(P, M)\,dv\right) + c. \quad (15)$$

The solution of the Neumann problem in the case of a plane domain D is determined by the formula

$$f(P) = -\frac{1}{2\pi}\oint\limits_{\gamma} \beta(N)G(P, N)\,dl - \frac{1}{2\pi}\iint\limits_{D} g(M)G(P, M)\,d\sigma + c, \quad (16)$$

the Green function $G(P, M)$ satisfying on the contour γ the condition

$$\frac{dG}{dn}\Big|_{\gamma} = -\frac{2\pi}{L(\gamma)},$$

where $L(\gamma)$ is the length of the contour γ.

The derivation of formula (16) is analogous to that of formula (13), and we leave it to the reader.

4. Solution of the Dirichlet Problem for a Ball and a Circle. The problem of finding Green's function is sufficiently complicated and requires cumbersome calculations even for simple domains. There are various methods of constructing Green's function. One of them is the method of symmetry. As an example, let us consider the construction of Green's function for the Dirichlet problem in the case of the ball.

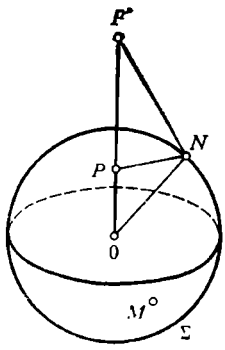


Fig. 35

Let us find Green's function (5) of the Dirichlet problem

$$\Delta f = g, \quad f|_{\Sigma} = \alpha(N)$$

for the case when the domain Ω is a ball of radius R with centre at the point O (Fig. 35). To this end, it is required to construct the harmonic function $\Phi(P, M)$ from the boundary conditions

$$\Phi(P, M)|_{\Sigma} = \Phi(P, N) = -\frac{1}{r_{NP}}, \tag{17}$$

where P is a fixed point, and N is an arbitrary boundary point. Let the point P be situated at a distance $OP = \rho$ from the centre of the sphere. We denote by P' the point symmetric to P with respect to the sphere (see Fig. 35). It lies on one and the same radius with the point P and satisfies the condition $\overline{OP}\cdot\overline{OP'} = R^2$; since $\overline{OP'} = \lambda\overline{OP}$, we get $\lambda\overline{OP}^2 = R^2$, $\lambda = R^2/\overline{OP}^2 = R^2/\rho^2$ and $\overline{OP} = \frac{R^2}{\rho^2}\overline{OP}$. Let M be an arbitrary point lying inside the sphere Σ, then the function $1/r_{P'M}$ is harmonic inside the sphere. We will seek the function $\Phi(P, M)$ in the form $c/r_{P'M}$, where $c = \text{const}$. We choose c so as to satisfy the boundary conditions (17):

$$\frac{c}{r_{P'M}}\Big|_{\Sigma} = \frac{c}{r_{P'N}} = -\frac{1}{r_{PN}}, \quad c < 0.$$

Hence we derive the relationship $r^2_{P'N} = c^2 r^2_{PN}$, and rewrite it in the form $\overline{P'N}^2 = c^2 \overline{PN}^2$. Bearing in mind that $\overline{P'N} = \overline{ON} - \overline{OP'}$, we obtain for a scalar square the following representation:

$$\overline{P'N}^2 = (\overline{ON} - \overline{OP'})^2 = \overline{ON}^2 - 2\overline{ON}\cdot\overline{OP'} + \overline{OP'}^2$$
$$= \frac{R^2}{\rho^2}(R^2 - 2\overline{ON}\cdot\overline{OP} + \rho^2).$$

The expression in the parentheses is the squared length of the vector $\overline{PN}$. Indeed

$$\overline{PN}^2 = (\overline{ON} - \overline{OP})^2 = R^2 - 2\overline{ON}\cdot\overline{OP} + \rho^2.$$

Thus, $\overline{P'N}^2 = c^2\overline{PN}^2 - \frac{R^2}{\rho^2}\overline{PN}^2$. $c < 0$, and consequently, $c = -\frac{R}{\rho}$, $|\overline{P'N}| = \frac{R}{\rho}|\overline{PN}|$. Then the harmonic function has the form $\Phi(P, M) = -\frac{R}{\rho}\frac{1}{r_{P'M}}$. Substituting this function into formula (5), we find Green's function:

$$G(P, M) = \frac{1}{r_{PM}} - \frac{R}{\rho}\frac{1}{r_{P'M}}.$$

Now, let us take advantage of formula (8) determining the solution of the Dirichlet problem. For this purpose, we find the value of the normal derivative dG/dn on the sphere:

$$\left.\frac{dG}{dn}\right|_\Sigma = \left.\frac{dG}{dr}\right|_\Sigma = \operatorname{grad} G\cdot\frac{\overline{ON}}{R}.$$

The gradient is directed along the normal moving up towards the level surface; to compute it, let us make use of formula (9) from Sec. 1.2

$$\operatorname{grad} G(P, N) \quad \nabla\left(\frac{1}{r_{PN}} - \frac{R}{\rho}\frac{1}{r_{P'N}}\right) \quad - \frac{\overline{PN}}{r^3_{PN}} + \frac{R}{\rho}\frac{\overline{P'N}}{r^3_{P'N}}.$$

Let us transform this expression, taking into account that $r_{P'N} = \frac{R}{\rho}r_{PN}$, then

$$\operatorname{grad} G(P, N) = \frac{1}{r^3_{PN}}\left(\frac{\rho^3}{R^3}\overline{P'N} - \overline{PN}\right)$$
$$= \frac{1}{r^3_{PN}}\left(\frac{\rho^3}{R^3}\overline{ON} - \overline{OP} - \overline{ON} + \overline{OP}\right) = \frac{\rho^2 - R^2}{R^3\cdot r^3_{PN}}\overline{ON}.$$

Consequently, for the normal derivative we obtain the expression

$$\frac{dG}{dn}\Big|_{\Sigma} = \operatorname{grad} G \cdot \mathbf{n}^0 \,|_{\Sigma} = \frac{\rho^2 - R^2}{R} \frac{1}{r_{PN}^3}.$$

Substituting G and $\frac{dG}{dn}\Big|_{\Sigma}$ into formula (8), we get the solution of the Dirichlet problem of Poisson's equation for a ball of radius R in integral form:

$$f(P) = \frac{R^2 - \rho^2}{4\pi R} \oiint_{\Sigma} \frac{\alpha(N)}{r_{PN}^3} \, d\sigma$$

$$- \frac{1}{4\pi} \iiint_{\Omega} \left(\frac{1}{r_{PM}} - \frac{R}{\rho} \frac{1}{r_{P'M}} \right) g(M) \, dv. \qquad (18)$$

In particular, for Laplace's equation the solution is represented by the surface integral

$$f(P) = \frac{R^2 - \rho^2}{4\pi R} \oiint_{\Sigma} \frac{\alpha(N)}{r_{PN}^3} \, d\sigma, \qquad (19)$$

which is called *Potsson's integral.*

The solution of the Dirichlet problem for Poisson's and Laplace's equations in integral form in the case of a ball is given by formulas (18) and (19) respectively. The integrals in formulas (18) and (19) are found, as a rule, by applying approximate methods.

For a circle the Dirichlet problem is solved just in the same manner as for a ball.

As an exercise, we want the reader to show that Green's function in the case of a circle has the form

$$G(P, M) = \ln \frac{1}{r_{PM}} - \ln \left(\frac{R}{\rho} \frac{1}{r_{P'M}} \right),$$

where P' is a point symmetric to the point P with respect to the circle of radius R, and the solution of the Dirichlet

problem for Poisson's equation is given by the formula

$$f(P)=\frac{R^2-\rho^2}{2\pi R}\oint_\gamma \frac{\alpha(N)}{r_{PN}^2}\,dl$$
$$-\frac{1}{2\pi}\iint_D g(M)\left(\ln\frac{1}{r_{PM}}-\ln\frac{R}{\rho r_{PM}}\right)d\sigma. \quad (20)$$

In particular, for Laplace's equation the solution is written with the aid of the line integral:

$$f(P)=\frac{R^2-\rho^2}{2\pi R}\oint_\gamma \frac{\alpha(N)}{r_{PN}^2}\,dl. \quad (21)$$

Integral (21), the same as (19), is called *Poisson's integral.*

Let us reduce the line integral (21) to a definite integral. Let $P(\rho, \gamma)$ and $N(R, \varphi)$ be the polar coordinates, then

$$r_{PN}^2 = R^2 + \rho^2 - 2R\rho\cos(\varphi-\gamma), \quad dl = R\,d\varphi$$

and Poisson's integral (21) takes the form

$$f(P)=\frac{R^2-\rho^2}{2\pi}\int_0^{2\pi}\frac{\alpha(R,\varphi)}{R^2+\rho^2-2R\rho\cos(\varphi-\gamma)}\,d\varphi. \quad (22)$$

Sec. 2.5. DERIVING CERTAIN EQUATIONS OF MATHEMATICAL PHYSICS

1. Continuity Equation. As an application of the vector field theory in physics, we are going to give the derivation of one of the equations of the motion of fluid, namely, the continuity equation. We will regard the motion of gas as the motion of a compressible liquid whose density $\rho = \rho(M, t)$ is a function of point M and time t. The motion of a liquid can be characterized by specifying the velocity field $\mathbf{v} = \mathbf{v}(M, t)$. For any motion of liquid the functions $\mathbf{v}$ and ρ will be interrelated by the equation which is called the *continuity equation.*

When deriving this equation, we use the method in which the change in the mass of the liquid found inside an arbitrarily taken surface Σ bounding the domain Ω is computed in two different ways. The amount of the liquid found in a given

volume at a given instant of time is computed with the aid of the triple integral $\iiint_{\Omega} \rho(M, t)\, d\tau$. During the time Δt the amount of the liquid Q in Ω will change by the quantity

$$\Delta Q = Q(t+\Delta t) - Q(t) = \iiint_{\Omega} \rho(M, t+\Delta t)\, d\tau - \iiint_{\Omega} \rho(M, t)\, d\tau = \iiint_{\Omega} \frac{\Delta_t \rho(M, t)}{\Delta t} \Delta t\, d\tau. \quad (1)$$

Let us now determine the change in the amount of the liquid occupying the same volume for the same interval of time in another way. The change in the amount of the liquid in the volume Ω thus determined may occur only owing to the fact that some amount of the liquid passed across the surface Σ bounding this volume. The amount of the liquid flown into Ω across the surface Σ per unit time is equal to the flux of the field $\rho\mathbf{v}$ which is expressed by the integral

$$-\iint_{\Sigma} \rho\mathbf{v} \cdot \mathbf{n}^0\, d\sigma.$$

Here, the minus sign indicates that the liquid flows in. During the time Δt the amount of the liquid passing across the surface Σ is computed by the formula

$$\Delta Q = \iint_{\Sigma} \rho\mathbf{v} \cdot \mathbf{n}^0\, d\sigma\, \Delta t.$$

Applying the Ostrogradsky-Gauss formula, we may write

$$\Delta Q = -\iint_{\Sigma} \rho\mathbf{v} \cdot \mathbf{n}^0\, d\sigma\, \Delta t = -\iiint_{\Omega} \operatorname{div}(\rho\mathbf{v})\, d\tau\, \Delta t. \quad (2)$$

Equating the expressions (1) and (2) obtained for ΔQ, we get the equality

$$\iiint_{\Omega} \left(\frac{\Delta_t \rho}{\Delta t} + \operatorname{div}(\rho\mathbf{v}) \right) d\tau\, \Delta t = 0$$

which holds true for any domain Ω and small Δt. By virtue of the arbitrariness of the domain Ω, the integrand must be identically equal to zero, which in the limit, as $\Delta t \to 0$, turns into the equality

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0. \tag{3}$$

This is just the desired continuity equation.

Equation (3) can be given another form frequently used in applications. Let us introduce the concept of the total derivative. Studying the change in the function $\rho(M, t)$ during a certain time interval Δt, we may proceed in two different ways: namely, to consider the change in ρ at a given place or to consider it for a given particle. The change in ρ at a given place is characterized by the partial derivative

$$\frac{\partial \rho}{\partial t} = \lim_{\Delta t \to 0} \frac{\rho(M, t+\Delta t) - \rho(M, t)}{\Delta t}, \tag{4}$$

in computation of which the radius vector of the point M is regarded as constant.

Let the particle M move with velocity $\mathbf{v}$ and, during the time Δt, let it displace into the point M'. The variation of ρ for the given particle during the time Δt is characterized by the total (or individual or substantial) derivative

$$\frac{d\rho}{dt} = \lim_{\Delta t \to 0} \frac{\rho(M', t+\Delta t) - \rho(M', t)}{\Delta t}. \tag{5}$$

Setting $M = M(t)$, we can find the total derivative by the formula

$$\frac{d\rho}{dt} = \frac{d\rho}{ds} \frac{ds}{dt} + \frac{\partial \rho}{\partial t},$$

where $d\rho/ds$ is the derivative with respect to the arc length of the curve. As is known, ds/dt represents the velocity at the point M, therefore we get

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \frac{d\rho}{ds} v = \frac{\partial \rho}{\partial t} + \operatorname{grad} \rho \cdot \mathbf{s}^0 v,$$

or

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \operatorname{grad} \rho. \tag{6}$$

Let us transform the continuity equation (3). We have

$$\operatorname{div}(\rho \mathbf{v}) = \nabla(\rho \mathbf{v}) = \nabla\rho \mathbf{v} + \rho \nabla \mathbf{v} = \mathbf{v} \cdot \operatorname{grad} \rho + \rho \operatorname{div} \mathbf{v},$$

therefore equation (3) takes the form

$$\frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \operatorname{grad} \rho + \rho \operatorname{div} \mathbf{v} = 0.$$

Using equality (6), we get the continuity equation in a different form:

$$\frac{d\rho}{dt} + \rho \operatorname{div} \mathbf{v} = 0. \tag{7}$$

Consider the particular case of an incompressible, but, possibly, nonhomogeneous liquid. In this case the density of each particle of the liquid remains unchanged and, consequently, by the definition of the total derivative, $d\rho/dt = 0$.

The continuity equation for an incompressible liquid has the form

$$\operatorname{div} \mathbf{v} = 0. \tag{8}$$

Thus, in the case of an incompressible liquid, the velocity vector is a solenoidal vector. And if, in addition, the velocity field is potential, that is, $\mathbf{v} = \operatorname{grad} f$, then the continuity equation (6) turns into Laplace's equation $\Delta f = 0$. Thus, the velocity potential in the potential motion of an incompressible liquid satisfies Laplace's equation.

2. Heat Equation. We shall hold that the thermal state of a certain body is known if for every point M of the body its temperature $T = T(M, t)$ is known at any instant of time t. To derive the heat equation, let us consider inside the body an arbitrary volume Ω bounded by the surface Σ and compute in two different ways the change in the quantity of heat contained in the volume Ω.

Regarding the medium as isotropic, let us denote the density of the body by $\rho = \rho(M)$, its heat capacity by $c = c(M)$, and the thermal conductivity by $k = k(M)$. For a nonhomogeneous body the quantities ρ, c, and k are functions of the point M, while for a homogeneous body these quantities are constants. The intensity of heat sources at the point M at time t will be denoted by $f_1(M, t)$. Let us now compute the heat balance in the volume Ω during the time interval $(t, t + \Delta t)$. According to the Fourier law, the quantity of heat entering the volume Ω through the

surface Σ during the time Δt is proportional to the flux of temperature T across the surface Σ, that is,

$$Q_1 = \iint\limits_{\Sigma} k \frac{dT}{dn} d\sigma \Delta t = \iint\limits_{\Sigma} k \operatorname{grad} T \cdot \mathbf{n}^0 d\sigma \Delta t.$$

Applying the Ostrogradsky-Gauss formula, we may write

$$Q_1 = \int\int\int\limits_{\Omega} \operatorname{div}(k \operatorname{grad} T)\, dv \Delta t.$$

Besides, the thermal sources situated in the volume Ω itself during the time interval Δt generate the quantity of heat

$$Q_2 = \int\int\int\limits_{\Omega} f_1(M, t)\, dv \Delta t.$$

The quantity of heat Q_3 required to change the temperature of the body having the volume Ω by the quantity $\Delta T = T(M, t + \Delta t) - T(M, t) \approx (\partial T/\partial t) \Delta t$ is expressed with the aid of the integral

$$Q_3 = \int\int\int\limits_{\Omega} c\rho \frac{\partial T}{\partial t}\, dv \Delta t.$$

Applying the condition of heat balance $Q_3 = Q_1 + Q_2$, we get the equality

$$\int\int\int\limits_{\Omega} \left(\operatorname{div}(k \operatorname{grad} T) + f_1 - c\rho \frac{\partial T}{\partial t} \right) dv \Delta t = 0.$$

Since this equality takes place for an arbitrary volume Ω, the integrand is identically equal to zero, that is,

$$c\rho \frac{\partial T}{\partial t} = \operatorname{div}(k \operatorname{grad} T) + f_1. \tag{9}$$

Equation (9) is called the *heat equation*. In the case of a homogeneous body, the quantities c, ρ, and k are constants, therefore, introducing the notation: $a^2 = k/c\rho$ and $f = f_1/c\rho$, we rewrite the heat equation in the form

$$\frac{\partial T}{\partial t} = a^2 \Delta T + f. \tag{10}$$

For a complete description of the thermal state of a body,

it is necessary, in addition to equation (9) or (10), to specify the initial distribution of temperature and boundary conditions, i.e. the boundary-value problems for these equations.

Boundary-value problems for the heat equation are set in the following way. Find the solution of equation (9) or (10) satisfying the initial condition $T(M, 0) = \eta(M)$ and one of the boundary conditions: (1) on the boundary Σ the given temperature distribution $T|_\Sigma = \alpha(N)$ is maintained; (2) on the boundary Σ the heat flow $k \frac{dT}{dn}\Big|_\Sigma = \beta(N)$ is given; (3) on the boundary Σ heat exchange takes place according to Newton's law

$$\left(k \frac{dT}{dn} + \lambda (T - T_0)\right)\Big|_\Sigma = 0,$$

where λ is the heat-exchange coefficient, and T_0 is the ambient temperature.

Depending on boundary conditions, the problems under consideration are called the boundary-value problems of the first, second, and third kind.

Remark. Equation (8) was obtained for a spatial domain. It will remain true for a plane domain and for a rod as well. In Cartesian coordinates, equation (8) for the spatial case has the form

$$\frac{\partial T}{\partial t} = a^2 \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right) + f(x, y, z, t).$$

For a plane domain, we obtain a two-dimensional heat equation which is written in the form

$$\frac{\partial T}{\partial t} = a^2 \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + f(x, y, t),$$

and a one-dimensional heat equation in the form

$$\frac{\partial T}{\partial t} = a^2 \frac{\partial^2 T}{\partial x^2} + f(x, t).$$

If the function of temperature T does not depend on time, then we speak of a stationary heat-conduction problem. It is just a stationary process that is characterized by the function $T = T(M)$ which depends only on the position of a point and is independent of time, therefore $\partial T/\partial t = 0$. The stationary temperature distribution inside a homoge-

neous body is described by Poisson's equation $\Delta T = \psi(M)$, where $\psi(M) = -f(M)/a^2$. If heat sources are absent, then Poisson's equation turns into Laplace's equation $\Delta T = 0$. Naturally, there are no initial conditions for stationary processes. For a complete description of a process only boundary conditions are given. Boundary-value problems of the first and second kind turn into the Dirichlet and Neumann problems.

Example 1°. Determine the stationary distribution of temperature inside a spherical layer $a < r < b$ if the sphere $r = a$ is maintained at a temperature T_1, and the sphere $r = b$ at a temperature T_2.

It is more convenient to solve this problem in spherical coordinates. For the reasons of symmetry, the temperature $T = T(r)$ depends only on the radius r and is independent of the variables θ and φ. The stationary distribution of temperature satisfies Laplace's equation $\Delta T = 0$. For the function $T(r)$ the Laplacian in the spherical coordinate system takes the form

$$\Delta T = \frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dT}{dr}\right).$$

Thus, we arrive at the equation

$$\frac{d}{dt}\left(r^2\frac{dT}{dr}\right) = 0,$$

from which we find $T = c_1/r + c_2$. The constants c_1 and c_2 are determined from the boundary conditions

$$T_1 = c_1/a + c_2, \quad T_2 = c_1/b + c_2.$$

Solving this system, we find

$$c_1 = \frac{ab(T_1 - T_2)}{b-a} \quad \text{and} \quad c_2 = \frac{bT_2 - aT_1}{b-a}.$$

Consequently, the desired solution has the form

$$T = \frac{ab(T_1 - T_2)}{b-a}\,\frac{1}{r} + \frac{bT_2 - aT_1}{b-a}.$$

Thus, in the example under consideration the temperature on each sphere of radius r is constant, and it changes according to the hyperbolic law when passing from one sphere to another.

CHAPTER 3

Certain Concepts of Functional Analysis

Sec. 3.1. STATEMENT OF PROBLEMS. HÖLDER'S AND MINKOWSKI'S INEQUALITIES

1. Fundamental Problems. As an independent mathematical discipline, functional analysis appeared at the beginning of the twentieth century. Nevertheless, developing in an exclusively rapid way, to the present days functional analysis has become an important branch of mathematics having numerous applications in both pure mathematical branches and in so-called applied mathematics. Functional analysis has formed as a result of generalization of various concepts and methods used in "elder" mathematical disciplines. This generalization was achieved by passing to a higher step of mathematical abstraction, the latter being characteristic of the methods of modern mathematics. Considering various mathematical, physical, and engineering problems from a more general, abstract point of view enables us rather frequently to reveal their laws and relationships to a better effect, to bring to light the common features inherent in the problems, similar by the methods of their solution, but different by their concrete content. Today, it is difficult to imagine the solution of any serious question from the field of differential equations of mathematical physics, approximate calculations, and a number of other problems without applying the methods of functional analysis.

The questions under consideration may belong to different mathematical problems having a definite meaning, but when we abstract their concrete content, these questions may be combined mathematically in one or two problems. Indeed, let X be a set of elements x (they may be numbers, functions, vectors, matrices, etc.), and let T be some map-

ping of the set X into itself, i.e. $Tx \in X$. Then the equation

$$Tx = \theta, \tag{1}$$

where θ is a zero element of the set X, means that we seek for the element $x_0 \in X$ which is transformed into a zero element by T. To such a problem we reduce, for instance, the solution of the system of linear equations

$$\sum_{k=1}^{n} a_{jk}x_k = r_j, \quad j = 1, 2, \ldots, n. \tag{2}$$

Indeed, setting $Tx = Ax - r$, where

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad r = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix},$$

we will reduce the solution of system (2) to finding an element x_0 such that $Tx_0 = \theta$, where $\theta = (0, 0, \ldots, 0)^T$.

Analogously, the problem of finding the solution of the system of linear differential equations

$$\frac{dy_j}{dt} = \sum_{\nu=1}^{n} a_{j\nu}y_\nu, \quad j = 1, 2, \ldots, n,$$

or the problem of finding the eigenvalues of the Sturm-Liouville boundary-value problem leads to solving the equation

$$Tx = \lambda x, \tag{3}$$

where λ is a real number, that is, to finding the element x which is transformed into a collinear element by T.

Considered with respect to equations (1) and (3) may be a number of problems, the following being the fundamental problems:

I. *Solution existence problem.*

II. *Solution uniqueness problem.*

III. *Problem of the choice of a method for an exact or approximate solution.*

IV. *Solution stability problem.*

V. *Problem of error estimate in an approximate solution.* Of course, it is clear that the solution of Problems II-V has meaning only when the solution existence problem is solved in a positive sense. It is also clear that the method of solving any of the listed problems in the general case is applicable to a number of particular cases.

As it was noted above, the elements of the sets X, Y, ... may have different nature. We shall consider abstract mathematical sets, for instance: E_1 which is the set of all real numbers; $[a, b]$—the set of real numbers t satisfying the condition $a \leqslant t \leqslant b$; the set of square matrices of order n; the set of linearly independent solutions of a homogeneous linear differential equation of order n; the set of states of some cybernetic system, and so on. The same as on the sets of real numbers, on arbitrary sets we also introduce the notion of functional dependence.

Let there be given two sets X and Y. We say that on the set X an operator f is specified with values in Y if each element $x \in X$ is associated with the unique element $y \in Y$ according to a definite rule. We shall denote this correspondence as follows: $y = f(x)$ or $y = fx$. We shall also say that we have the mapping f of the set X into the set Y, that is, $fX \subset Y$.

There is a variety of mappings, for example: (a) a mirror image of points in space with respect to a plane; (b) a projection of points in space on some plane or straight line; (c) a conformal mapping of the domain D of the complex plane on the domain G; (d) a mapping of the set of square integrable on $[a, b]$ functions into the set of sequences of the coefficients of the expansions of these functions with respect to a system orthonormal on $[a, b]$, and so forth.

If $Y = E_1$, that is, the set X is mapped into the set E_1 of real numbers, then the operator f is called the *functional.* In other words, the functional given on the set X is defined as the mapping of the set X into the set E_1 of real numbers.

The study of the properties of functions of a real variable is substantially based on the concept of limit in the set of real numbers. Therefore, it is quite natural to introduce also the concept of limit into the sets under consideration. This leads to a new notion of *space* which is understood as some set with the notion of limit introduced. In general,

the concept of space in science may have various sense. Thus, in philosophy space is one of the forms of existence of matter, in geometry space is understood as the space of three dimensions E_3 with a certain system of axioms. Geometry mainly studies three-dimensional forms. In applications, we often deal with variables whose number n exceeds three. Thus, to the coordinates of points we add time, velocity, acceleration, and a number of other variables. This led to the necessity of abstract generalization of the three-dimensional space E_3 for the needs of n-dimensional and infinite dimensional spaces and to the study of the properties of such spaces. For their introduction and analysis, let us prove a number of inequalities which play an important role in many applications.

2. Auxiliary Inequalities. Let us first establish two lemmas.

Lemma 1. *Let us have* $u > 0$, $v > 0$, $p > 1$, *and* q *is conjugate to* p, *i.e.* $(1/p) + (1/q) = 1$; *then*

$$uv \leqslant \frac{u^p}{p} + \frac{v^q}{q}. \tag{4}$$

Consider the function $y = x^\alpha$ for $\alpha > 0$. Since y increases monotonically, there exists the inverse function $x = y^{1/\alpha}$.

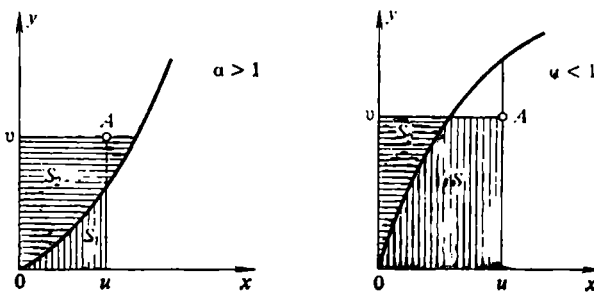


Fig. 36

Let us take the points $x = u$ and $y = v$ on the x- and y-axis, respectively, and let us consider the areas S_1 and S_2 (Fig. 36). It is clear that the sum of the areas S_1 and S_2 is no less than the area of the rectangle $AuOv$ equal to uv.

But

$$S_1 = \int_0^u x^\alpha \, dx = \frac{u^{\alpha+1}}{\alpha+1},$$

$$S_2 = \int_0^v y^{1/\alpha} \, dy = \frac{v^{1/\alpha+1}}{1/\alpha+1}.$$

therefore we get the inequality

$$uv \leqslant \frac{u^{\alpha+1}}{\alpha+1} + \frac{v^{1/\alpha+1}}{1/\alpha+1}.$$

Setting here $\alpha + 1 = p$ and $1/\alpha + 1 = q$, we come to inequality (4). Note that the equality sign in (4) takes place only if $v = u^{\alpha-1} = u^{p-1}$.

Lemma 2. *For any values of a and b the following inequality holds*:

$$|a + b|^p \leqslant 2^p (|a|^p + |b|^p), \quad p \geqslant 1. \tag{5}$$

In fact, if $|a| \leqslant |b|$, then we have $|a + b| \leqslant 2|b|$, i.e.

$$|a + b|^p \leqslant 2^p |b|^p \leqslant 2^p (|a|^p + |b|^p).$$

Similarly, for $|a| > |b|$ we get the inequality $|a + b| < 2|a|$, whence it follows that

$$|a + b|^p < 2^p |a|^p \leqslant 2^p (|a|^p + |b|^p),$$

that is, inequality (5) is fulfilled for any a and b.

3. Hölder's Inequalities for Integrals and Sums. The functions considered in applications have, as a rule, no more than a finite number of points of discontinuity or a denumerable set having a finite number of limit points. Such functions are integrable in the sense of the Riemann integral introduced in the course of mathematical analysis. But functional analysis uses a more general notion of the Lebesgue integral. Bearing in mind that a function which has a Riemann integral necessarily has a Lebesgue integral (but not conversely), we shall not give the definition of the Lebesgue integral, although individual notions (for instance, the completeness of the space of functions that are integrable in some power) are valid only if the classes of the functions under consideration are appropriately completed. We would like to note here that the computations given later for integrals are also true for the general Lebesgue integral.

Theorem 1. *Let the numbers p and q satisfy the condition $(1/p) + (1/q) = 1$ and the functions $x(t)$ and $y(t)$ defined on $[a, b]$ are such that the integrals*

$$\int_a^b |x(t)|^p \, dt \quad \text{and} \quad \int_a^b |y(t)|^q \, dt$$

exist and are different from zero. Then the product $|x(t)\, y(t)|$ is also integrable on $[a, b]$ and we have Hölder's inequality

$$\int_a^b |x(t)\, y(t)| \, dt \leqslant \left(\int_a^b |x(t)|^p \, dt\right)^{1/p} \left(\int_a^b |y(t)|^q \, dt\right)^{1/q}. \quad (6)$$

We set

$$u = \frac{|x(t)|}{\left(\int_a^b |x(t)|^p \, dt\right)^{1/p}} \quad \text{and} \quad v = \frac{|y(t)|}{\left(\int_a^b |y(t)|^q \, dt\right)^{1/q}}.$$

Applying inequality (4), we get the relationship

$$uv = \frac{|x(t)\, y(t)|}{\left(\int_a^b |x(t)|^p \, dt\right)^{1/p} \left(\int_a^b |y(t)|^q \, dt\right)^{1/q}} \leqslant \frac{|x(t)|^p}{p \int_a^b |x(t)|^p \, dt} + \frac{|y(t)|^q}{q \int_a^b |y(t)|^q \, dt}.$$

By the hypothesis, the right-hand side is integrable on $[a, b]$, therefore the left-hand side is also integrable, and we get the estimate:

$$\frac{\int_a^b |x(t)\, y(t)| \, dt}{\left(\int_a^b |x(t)|^p \, dt\right)^{1/p} \left(\int_a^b |y(t)|^q \, dt\right)^{1/q}} \leqslant \frac{\int_a^b |x(t)|^p \, dt}{p \int_a^b |x(t)|^p \, dt} + \frac{\int_a^b |y(t)|^q \, dt}{q \int_a^b |y(t)|^q \, dt} = \frac{1}{p} + \frac{1}{q} = 1.$$

Hence there follows Hölder's inequality (6).

If $p = q = 2$, then from inequality (6) it is possible to derive the *Cauchy-Buniakowski inequality* (proved in the first part of the book, Sec. 7.3) which is also called *Schwarz's inequality*:

$$\int_a^b |x(t)\,y(t)|\,dt \leqslant \sqrt{\int_a^b |x(t)|^2\,dt \int_a^b |y(t)|^2\,dt}. \quad (7)$$

Theorem 2. *Let* $(1/p) + (1/q) = 1$ *and the sequences* $\{u_k\}$ *and* $\{v_k\}$ *be such that the series*

$$\sum_{k=1}^{\infty} |u_k|^p \quad \textit{and} \quad \sum_{k=1}^{\infty} |v_k|^q$$

converge. Then the series $\sum_{k=1}^{\infty} |u_k v_k|$ *also converges, and Hölder's inequality holds true*

$$\sum_{k=1}^{\infty} |u_k v_k| \leqslant \Big(\sum_{k=1}^{\infty} |u_k|^p\Big)^{1/p} \Big(\sum_{k=1}^{\infty} |v_k|^q\Big)^{1/q}. \quad (8)$$

Let us set

$$u = \frac{|u_k|}{\Big(\sum\limits_{\nu=1}^{\infty} |u_\nu|^p\Big)^{1/p}} \quad \text{and} \quad v = \frac{|v_k|}{\Big(\sum\limits_{\nu=1}^{\infty} |v_\nu|^q\Big)^{1/q}}.$$

Applying inequality (4), we find the estimate

$$uv = \frac{|u_k v_k|}{\Big(\sum\limits_{\nu=1}^{\infty} |u_\nu|^p\Big)^{1/p} \Big(\sum\limits_{\nu=1}^{\infty} |v_\nu|^q\Big)^{1/q}} \leqslant \frac{|u_k|^p}{p \sum\limits_{\nu=1}^{\infty} |u_\nu|^p} + \frac{|v_k|^q}{q \sum\limits_{\nu=1}^{\infty} |v_\nu|^q}. \quad (9)$$

The right-hand side of this inequality is a term of a convergent numerical series, and, consequently, the series

$\sum_{k=1}^{\infty} |u_k v_k|$ converges. Therefore, summing together inequalities (9) with respect to k from 1 to ∞, we obtain Hölder's inequality for sums (8).

Setting in (8) $p = q = 2$, we obtain the inequality

$$\sum_{k=1}^{\infty} |u_k v_k| \leqslant \sqrt{\sum_{k=1}^{\infty} |u_k|^2 \sum_{k=1}^{\infty} |v_k|^2} \tag{10}$$

called *Cauchy's inequality.*

4. Minkowski's Inequalities for Integrals and Sums. The inequalities established in the following theorems are widely used in functional analysis, theory of orthogonal series, and in a number of other fields of mathematics.

Theorem 3. *Let the functions $x(t)$ and $y(t)$ defined on $[a, b]$ be such that the integrals*

$$\int_a^b |x(t)|^p \, dt < \infty, \quad \int_a^b |y(t)|^p \, dt < \infty, \quad p \geqslant 1, \tag{11}$$

are existent and finite. Then the function $|x(t) + y(t)|^p$ is also integrable, and we have Minkowski's inequality

$$\left(\int_a^b |x(t) + y(t)|^p \, dt \right)^{1/p} \leqslant \left(\int_a^b |x(t)|^p \, dt \right)^{1/p} + \left(\int_a^b |y(t)|^p \, dt \right)^{1/p}.$$

Applying inequality (5), we obtain the estimate

$$|x(t) + y(t)|^p \leqslant 2^p [|x(t)|^p + |y(t)|^p],$$

whence it follows that the following integral is finite:

$$\int_a^b |x(t) + y(t)|^p \, dt < \infty.$$

Representing first this integral in the form

$$\int_a^b |x(t)+y(t)|^p\,dt = \int_a^b |x(t)+y(t)|^{p-1}\,|x(t)+y(t)|\,dt$$

$$\leqslant \int_a^b |x(t)+y(t)|^{p-1}\,|x(t)|\,dt$$

$$+ \int_a^b |x(t)+y(t)|^{p-1}\,|y(t)|\,dt, \quad (12)$$

we estimate it.

Putting $q = p/(p-1)$ and noting that

$$\int_a^b [|x(t)+y(t)|^{p-1}]^q\,dt = \int_a^b |x(t)+y(t)|^{\frac{(p-1)p}{p-1}}\,dt$$

$$= \int_a^b |x(t)+y(t)|^p\,dt < \infty,$$

we apply Hölder's inequality (6) to the right-hand side of inequality (12). Then we get the following estimate:

$$\int_a^b |x(t)+y(t)|^p\,dt \leqslant \left(\int_a^b |x(t)|^p\,dt\right)^{1/p}\left(\int_a^b |x(t)+y(t)|^{(p-1)q}\,dt\right)^{1/q}$$

$$+ \left(\int_a^b |y(t)|^p\,dt\right)^{1/p}\left(\int_a^b |x(t)+y(t)|^{(p-1)q}\,dt\right)^{1/q}$$

$$= \left(\int_a^b |x(t)+y(t)|^p\,dt\right)^{1/q}\left[\left(\int_a^b |x(t)|^p\,dt\right)^{1/p}\right.$$

$$\left. + \left(\int_a^b |y(t)|^p\,dt\right)^{1/p}\right].$$

Dividing both its sides by $\left(\int_a^b |x(t)+y(t)|^p\,dt\right)^{1/q}$ and taking into consideration that $1-(1/q) = 1/p$,

we obtain the inequality

$$\Big(\int_a^b |x(t)+y(t)|^p\,dt\Big)^{1/p} \leqslant \Big(\int_a^b |x(t)|^p\,dt\Big)^{1/p} + \Big(\int_a^b |y(t)|^p\,dt\Big)^{1/p}.$$

Theorem 4. *Let the sequences* $\{u_k\}$ *and* $\{v_k\}$ *be such that the series* $\sum_{k=1}^{\infty} |u_k|^p$ *and* $\sum_{k=1}^{\infty} |v_k|^p$, $p \geqslant 1$, *converge. Then the series* $\sum_{k=1}^{\infty} |u_k+v_k|^p$ *also converges, and we have Minkowski's inequality for the series*

$$\Big(\sum_{k=1}^{\infty} |u_k+v_k|^p\Big)^{1/p} \leqslant \Big(\sum_{k=1}^{\infty} |u_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{\infty} |v_k|^p\Big)^{1/p}. \quad (13)$$

From inequality (5) we have

$$|u_k + v_k|^p \leqslant 2^p (|u_k|^p + |v_k|^p),$$

wherefrom there follows the convergence of the series $\sum_{k=1}^{\infty} |u_k + v_k|^p$. As in proving Theorem 3, let us estimate this series using Hölder's inequality for sums:

$$\begin{aligned}\sum_{k=1}^{\infty} |u_k+v_k|^p &\leqslant \sum_{k=1}^{\infty} |u_k+v_k|^{p-1}|u_k| + \sum_{k=1}^{\infty} |u_k+v_k|^{p-1}|v_k| \\ &\leqslant \Big(\sum_{k=1}^{\infty} |u_k|^p\Big)^{1/p} \Big(\sum_{k=1}^{\infty} |u_k+v_k|^{(p-1)q}\Big)^{1/q} \\ &\quad + \Big(\sum_{k=1}^{\infty} |v_k|^p\Big)^{1/p} \Big(\sum_{k=1}^{\infty} |u_k+v_k|^{(p-1)q}\Big)^{1/q} \\ &= \Big(\sum_{k=1}^{\infty} |u_k+v_k|^p\Big)^{1/q} \Big[\Big(\sum_{k=1}^{\infty} |u_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{\infty} |v_k|^p\Big)^{1/p}\Big].\end{aligned}$$

Dividing now both sides of this inequality by $\big(\sum_{k=1}^{\infty} |u_k+v_k|^p\big)^{1/q}$, we obtain inequality (13).

Sec. 3.2. METRIC SPACES

1. Metric and Limit in a Metric Space. In mathematical analysis we encountered different types of convergence determined by different concepts of limit. Let us recall some of them:

(a) in the set of real numbers E_1 (or in the set of complex numbers K) the limit of the sequence $\{x_n\}$, i.e. $x = \lim\limits_{n\to\infty} x_n$, means that $\forall\, \varepsilon > 0\ \exists\, N = N(\varepsilon)$ such that

$$|x_n - x| < \varepsilon \quad \forall\, n > N(\varepsilon);$$

(b) in the set of functions continuous on the interval $[a, b]$ the uniform convergence of the sequence of functions $\{f_n(x)\}$ to the function $f(x)$ means that $\forall\, \varepsilon > 0\ \exists\, N = N(\varepsilon)$ such that

$$\max_{a\leqslant x\leqslant b} |f_n(x) - f(x)| < \varepsilon \quad \forall\, n > N(\varepsilon);$$

(c) in the set of functions square integrable on $[a, b]$ the convergence in the mean of the sequence of functions $\{f_n(x)\}$ to the function $f(x)$ means the convergence to zero of the sequence of integrals

$$I_n = \int_a^b [f_n(x) - f(x)]^2\, dx;$$

(d) in the set of k-dimensional vectors the convergence of the sequence of vectors $\{x^{(n)} = (x_1^{(n)}, \ldots, x_k^{(n)})\}$ to the vector $a = (a_1, \ldots, a_k)$ means either the smallness of the quantities

$$\sqrt{\sum_{\nu=1}^{k} (x_\nu^{(n)} - a_\nu)^2} < \varepsilon \quad \forall\, n > N(\varepsilon)$$

or the smallness of the quantities

$$\max_{1\leqslant\nu\leqslant n} |x_\nu^{(n)} - a_\nu| < \varepsilon \quad \forall\, n > N(\varepsilon).$$

Noting that a continuous on $[a, b]$ function is square integrable on this interval, from examples (b), (c), and (d) we conclude that the "proximity" even of the same elements

of a set may be regarded in different meanings, depending on the chosen measure of "proximity" between the elements, which is called the *distance between the elements*. And if the notion of the distance between its elements is established in a set, then we can introduce the notion of limit in this set and turn this set into a space. In connection with this, let us give the following definitions.

A set X is called a *metric space* if to each pair x, y of its elements there is associated a nonnegative real number $\rho_X(x, y)$ which satisfies the following conditions:

(1) $\rho_X(x, y) = 0$ if and only if $x = y$;

(2) $\rho_X(x, y) = \rho_X(y, x)$ (symmetry axiom);

(3) for any elements x, y, and z from the set X the relationship $\rho_X(x, z) \leqslant \rho_X(x, y) + \rho_X(y, z)$ (triangle axiom).

The quantity $\rho_X(x, y)$ is called the *distance*, or *metric for the space* X, and conditions (1)-(3) are said to be *axioms of the metric*.

The metric space obtained from the set X by introducing in it the metric $\rho_X(x, y)$ will be denoted by X_ρ or simply by X when no misunderstanding may occur. When analyzing concrete metric spaces, we shall also write $\rho(x, y)$ instead of $\rho_X(x, y)$. With the aid of the metric, we can introduce the notion of limit in the metric space X_ρ.

The element $x \in X_\rho$ is said to be the *limit of the sequence* $\{x_n\} \subset X_\rho$ if $\rho(x_n, x) \to 0$ as $n \to \infty$.

Convergent sequences of a metric space have the properties of convergent sequences of real numbers.

As an example, let us give the proof of the following property.

Theorem 1. *The convergent sequence* $\{x_n\}$ *of the metric space* X_ρ *cannot have two limits.*

Suppose $\lim\limits_{n\to\infty} x_n = x$ and $\lim\limits_{n\to\infty} x_n = y$, i.e. $\rho(x_n, x) \to 0$ and $\rho(x_n, y) \to 0$ as $n \to \infty$. Then, by virtue of the triangle axiom, the following inequality holds true:

$$\rho(x, y) \leqslant \rho(x, x_n) + \rho(x_n, y).$$

But, by virtue of our supposition, the right-hand side of this inequality can be made arbitrarily small for sufficiently large n. This means that $\rho(x, y) = 0$. Then, by the first axiom of metric, we conclude that $x = y$.

Let us give one more definition which we will need later, namely, the definition of the notion of r-neighbourhood: the *r-neighbourhood of the point a*, $a \in X_\rho$, denoted by $S(a, r)$ is defined as a ball of radius r with centre at the point a, that is, as the set of points $x \in X_\rho$ such that $\rho(x, a) < r$.

2. Euclidean Space. By analogy with the three-dimensional space E_3, where three-dimensional vectors $x = (x_1, x_2, x_3)$ are considered, let us bring into consideration the set of n-dimensional vectors $x = (x_1, x_2, \ldots, x_n)$, where x_ν, $\nu = 1, 2, \ldots, n$, are real numbers. Let us determine the distance between the two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in the following way:

$$\rho(x, y) = \sqrt{\sum_{\nu=1}^{n} (x_\nu - y_\nu)^2}. \tag{1}$$

The fulfilment of the first two properties of the metric for this distance is obvious, and the third property is readily obtained by applying Minkowski's inequality for finite sums. Indeed, if $z = (z_1, \ldots, z_n)$, then we have the inequality

$$\rho(x, z) = \sqrt{\sum_{\nu=1}^{n} (x_\nu - z_\nu)^2} = \sqrt{\sum_{\nu=1}^{n} [(x_\nu - y_\nu) + (y_\nu - z_\nu)]^2}$$
$$\leqslant \sqrt{\sum_{\nu=1}^{n} (x_\nu - y_\nu)^2} + \sqrt{\sum_{\nu=1}^{n} (y_\nu - z_\nu)^2},$$

and this is the triangle inequality

$$\rho(x, z) \leqslant \rho(x, y) + \rho(y, z).$$

Using metric (1), let us introduce the notion of the limit of the sequence $\{x^{(k)}\}$ in the set of n-dimensional vectors, assuming that $x = \lim\limits_{k \to \infty} x^{(k)}$ if only

$$\lim_{k \to \infty} \rho(x^{(k)}, x) = \lim_{k \to \infty} \sqrt{\sum_{\nu=1}^{n} (x_\nu^{(k)} - x_\nu)^2} = 0. \tag{2}$$

From the last relationship there follow the inequalities

$$\lim_{k \to \infty} x_\nu^{(k)} = x_\nu, \quad \nu = 1, 2, \ldots, n,$$

from which we conclude that the convergence of the sequence $\{x^{(k)}\}$ also means the convergence of the sequences of coordinates $\{x_\nu^{(k)}\}$, $\nu = 1, 2, \ldots, n$.

Thus, the set of n-dimensional vectors with the notions of metric (1) and limit (2) introduced in it becomes a metric space which is denoted by E_n and is called the *n-dimensional vector space* or *n-dimensional Euclidean space.*

The properties of the E_n space for $n > 3$ are similar to those of the three-dimensional space. Let us dwell on some of them.

The vectors $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, $1 \leqslant m \leqslant n$, are called *linearly independent* if the equality $\sum_{\nu=1}^{m} \alpha_\nu x^{(\nu)} = \theta$ is possible only when $\alpha_\nu = 0$ for all $\nu = 1, 2, \ldots, m$ (here, θ denotes the zero element of the space E_n, that is, $\theta = (0, 0, \ldots, 0)$).

A *scalar product of the vectors* $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ denoted by (x, y) is defined as the sum of pair products of like coordinates, that is,

$$(x, y) = \sum_{\nu=1}^{n} x_\nu y_\nu.$$

Any n linearly independent vectors of the space E_n form the *basis of* this *space.* The basis $e^{(1)}, \ldots, e^{(n)}$ is said to be *orthogonal* if $(e^{(\nu)}, e^{(\mu)}) = 0$ for $\nu \neq \mu$. And if

$$(e^{(\nu)}, e^{(\mu)}) = \delta_{\nu, \mu} = \begin{cases} 0 & \text{for} \quad \nu \neq \mu, \\ 1 & \text{for} \quad \nu = \mu, \end{cases}$$

then the basis is called *orthonormal.*

Theorem 2. *With the aid of any system of m linearly independent vectors* $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, $1 \leqslant m \leqslant n$, *of the space* E_n, *it is possible to construct a system of orthonormal vectors* $e^{(1)}, \ldots, e^{(m)}$.

Let us take advantage of the *general orthogonalization method* suggested by Erhard Schmidt in 1905. We shall construct the orthonormal system $\{e^{(\nu)}\}$ successively. First of all let us set

$$e^{(1)} = x^{(1)} / \sqrt{(x^{(1)}, x^{(1)})}.$$

Let us now choose the constants α_{21} and α_{22} not vanishing simultaneously so that for the vector $e^{(2)} = \alpha_{21} e^{(1)} +$

$+ \alpha_{22}x^{(2)}$ the following relationships are fulfilled:

$$(e^{(2)}, e^{(1)}) = \alpha_{21} + \alpha_{22}(x^{(2)}, e^{(1)}) = 0 \qquad (3)$$

and

$$(e^{(2)}, e^{(2)}) = \alpha_{21}^2 + \alpha_{22}^2 (x^{(2)}, x^{(2)}) + 2\alpha_{21}\alpha_{22}(x^{(2)}, e^{(1)}) = 1.$$

The possibility of the choice of nonzero quantities α_{21} and α_{22} in equality (3) follows from the linear independence of the vectors $x^{(2)}$ and $e^{(1)} = x^{(1)}/\sqrt{(x^{(1)}, x^{(1)})}$. And it is clear that $\alpha_{22} \neq 0$. Further, we choose three constants α_{31}, α_{32}, and α_{33}, which are not all zero, so that for the vector

$$e^{(3)} = \alpha_{31}e^{(1)} + \alpha_{32}e^{(2)} + \alpha_{33}x^{(3)}$$

the following relationships are fulfilled:

$$(e^{(1)}, e^{(3)}) = (e^{(2)}, e^{(3)}) = 0 \quad \text{and} \quad (e^{(3)}, e^{(3)}) = 1.$$

The possibility of the choice of such constants also follows from the linear independence of the vectors $x^{(3)}$, $e^{(2)}$, and $e^{(1)}$. Continuing this process step by step, we shall just obtain an orthonormal system of the vectors $e^{(1)}, e^{(2)}, \ldots, e^{(m)}$.

The transformation of the space E_n into itself determined by the operator specified with the aid of the square matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

of order n is called *affine*. It will be written in the form $y^T = Ax^T$, where

$$x^T = (x_1, \ldots, x_n)^T = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix},$$

$$y^T = (y_1, \ldots, y_n)^T = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, \quad y \in E_n.$$

Depending on the values of the elements $a_{\nu k}$, the matrix A can realize various mappings. For instance, the operator

$$I = \begin{pmatrix} 1 & 0 \dots 0 \\ 0 & 1 \dots 0 \\ \dots & \dots \\ 0 & 0 \dots 1 \end{pmatrix} = (\delta_{\nu k})\Big|_{\nu,\, k=1}^{n}$$

is a *unit operator*, that is, any element $x \in E$ is mapped by this operator into itself: $Ix^T = x^T I = x^T$;

the operator

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \dots 0 \\ 0 & \lambda_2 \dots 0 \\ \dots & \dots \\ 0 & 0 \dots \lambda_n \end{pmatrix}$$

stretches the coordinates of the vector $x = (x_1, \dots, x_n)$ and maps this vector into the vector $y^T = Ax^T = (\lambda_1 x_1, \dots, \lambda_n x_n)^T$;

the operator

$$P_\nu = \begin{pmatrix} 0 \dots 0 & 0 & 0 \dots 0 \\ \dots & \dots & \dots \\ 0 \dots 0 & 0 & 0 \dots 0 \\ 0 \dots 0 & 1 & 0 \dots 0 \\ 0 \dots 0 & 0 & 0 \dots 0 \\ \dots & \dots & \dots \\ 0 \dots 0 & 0 & 0 \dots 0 \end{pmatrix} \leftarrow \nu\text{th row}$$

$$\uparrow$$

$$\nu\text{th column}$$

projects the vector $x = (x_1, \dots, x_n)$ on the unit vector $e^{(\nu)}$, i.e.

$$y^T = P_\nu x^T = [0, \dots, 0, x_\nu, 0, \dots, 0]^T.$$

We could give a number of other specific mappings, but we are not going to draw reader's attention to them.

Remark. In the set of n-dimensional vectors $x = (x_1, \dots, x_n)$, the metric can be defined by a method different

from formula (1). For instance, the following metric is often considered:

$$\rho_1(x, y) = \max_{1 \leqslant k \leqslant n} |x_k - y_k|.$$

It is clear that the metric $\rho(x, y)$ defined by relationship (1) and the metric $\rho_1(x, y)$ are interrelated by the inequalities

$$\rho_1(x, y) \leqslant \rho(x, y) \leqslant \sqrt{n}\,\rho_1(x, y).$$

Let us agree to denote the n-dimensional metric space obtained with the aid of the metric $\rho_1(x, y)$ by E_n^*.

Note that the set of elements contained in the unit ball of the space E_n is put in the set of elements of the unit ball of the space E_n^*.

3. Space of Continuous Functions. Let X be the set of functions $x(t)$ continuous on the interval $[a, b]$. Let us set

$$\rho_C(x, y) = \max_{a \leqslant t \leqslant b} |x(t) - y(t)|. \tag{4}$$

For any functions $x(t)$, $y(t)$, and $z(t)$ from the set X, we have the inequality

$$|x(t) - y(t)| \leqslant |x(t) - z(t)| + |z(t) - y(t)|,$$

wherefrom there follows the relationship

$$\max_t |x(t) - y(t)| \leqslant \max_t |x(t) - z(t)| + \max_t |z(t) - y(t)|,$$

and this means that the metric $\rho_C(x, y)$ satisfies the triangle inequality:

$$\rho_C(x, y) \leqslant \rho_C(x, z) + \rho_C(z, y).$$

The rest of the axioms of the metric are obvious.

The limit of the sequence of functions $\{x_n(t)\}$ defined with the aid of metric (4) means that

$$\max_{a \leqslant t \leqslant b} |x_n(t) - x(t)| \to 0 \quad \text{as} \quad n \to \infty. \tag{5}$$

But the fulfilment of condition (5) implies the uniform convergence of the sequence of functions $\{x_n(t)\}$ to the function $x(t)$. Thus, denoting by $C[a, b]$ the *space of continuous on*

$[a, b]$ *functions* $x(t)$ with metric (4), we conclude that the convergence in the space $C[a, b]$ is a uniform convergence of the sequence of functions.

4. Spaces of *p*th Power Integrable Functions. In many applications, for the same sets of functions various concepts of "proximity" are considered. Thus, in addition to the maximum of the modulus of deviation of functions considered in the preceding subsection, the concepts of mean deviation, mean-square deviation, etc. are introduced. The set of the functions X under consideration can also be extended. Indeed, let X be the set of defined on $[a, b]$ functions $x(t)$ the pth power of which ($1 \leqslant p < \infty$) is absolutely integrable on $[a, b]$

$$\int_a^b |x(t)|^p \, dx < \infty, \quad 1 \leqslant p < \infty.$$

We shall regard the functions $x(t)$ and $y(t)$ in this set X as *equivalent* if they differ at most by the set of points having the measure zero. Without introducing the definition of the Lebesgue measure, we confine ourselves to the sets M of the *Jordan measure zero*. This means that for any $\varepsilon > 0$ the set M can be covered by a finite number of open intervals the sum of the lengths of which is less than ε. It is clear that a set from a finite number of points or a countable set with a finite number of limit points have the Jordan measure zero.

Now, let us introduce a metric in the set X. Namely, for any functions $x(t)$ and $y(t)$ from X we determine the distance between them by the relationship

$$\rho_p(x, y) = \left(\int_a^b |x(t) - y(t)|^p \, dt \right)^{1/p}. \tag{6}$$

The existence of the integral on the right follows from Theorem 3 of the preceding section, which also implies the **triangle inequality**

$$\rho_p(x, y) = \left(\int_a^b |[x(t) - z(t)] + [z(t) - y(t)]|^p \, dt \right)^{1/p}$$

$$\leqslant \rho_p(x, z) + \rho_p(z, y).$$

The symmetry axiom for metric (6) is obvious, and the identity axiom is understood in the sense of equivalence of functions on the sets whose Jordan measures coincide. The convergence of the sequence of functions $\{x_n(t)\}$ to the function $x(t)$ is determined with the aid of the relationship

$$\rho_p(x_n, x) = \left(\int_a^b | x_n(t) - x(t) |^p \, dt \right)^{1/p} \to 0 \quad \text{as} \quad n \to \infty$$

and the obtained space will be denoted by $L^p[a, b]$. The convergence in this space $L^p[a, b]$ will be called *the convergence in the mean* with power (with exponent) p.

Note that if the numbers p_1 and p_2 satisfy the inequality $p_1 > p_2 \geqslant 1$, then for the spaces corresponding to them the following embedding is valid:

$$L^{p_1}[a, b] \subset L^{p_2}[a, b].$$

The space $L^2[a, b]$ is called the *Hilbert functional space.*

5. Space of Convergent Sequences. Let $X = \{a\}$ be a set of convergent sequences $a = (a_1, \ldots, a_n, \ldots)$, that is, of such sequences that the limit $\lim\limits_{n \to \infty} a_n$ exists and is finite. The metric is introduced in X in the following way: if $x = (x_1, \ldots, x_n, \ldots) \in X$ and $y = (y_1, \ldots, y_n, \ldots) \in X$, then the distance between them is determined by the equality

$$\rho(x, y)_c = \sup_n | x_n - y_n |. \tag{7}$$

Remark. The maximum of the differences $| x_n - y_n |$ with respect to all n may even not exist, whereas the upper bound does exist. For instance, if the sequences are given with the aid of the formulas

$$x_n = \frac{n}{n+1}, \qquad y_n = \frac{1}{n+1}, \qquad n = 1, 2, \ldots,$$

then for their difference the following relationship is valid:

$$x_n - y_n = \frac{n-1}{n+1} \to 1 \quad \text{as} \quad n \to \infty,$$

i.e. $\sup\limits_n (x_n - y_n) = 1$, although $\max\limits_n (x_n - y_n)$ is not existent.

The fulfilment of the metric axioms for the distance introduced in (7) is obvious.

The set X with metric (7) introduced will be called the *space of convergent sequences* and will be denoted by c.

Let us clarify the meaning of convergence in this space c. Let the sequence $\{a^{(k)}\} = \{(a_1^{(k)}, \ldots, a_n^{(k)}, \ldots)\} \subset c$ converge in c to the element $a = (a_1, \ldots, a_n, \ldots)$. This means that

$$\lim_{k\to\infty} \rho\,(a^{(k)},\ a)_c = \lim_{k\to\infty} \sup_n |\ a_n^{(k)} - a_n\ | = 0,$$

i.e. $\forall\, \varepsilon > 0\ \exists\ K = K\,(\varepsilon)$ such that

$$\sup_n |\ a_n^{(k)} - a_n\ | < \varepsilon \quad \text{for all} \quad k < K\,(\varepsilon).$$

Hence it follows that $|\ a_n^{(k)} - a_n\ | < \varepsilon$ for all $n = 1, 2, \ldots$ if only $k > K\,(\varepsilon)$. But this is a uniform coordinate-wise convergence of the sequence $\{a^{(k)}\}$.

6. **Spaces of Sequences with Convergent Series.** Let $X = \{x\}$ be the set of sequences $x = (x_1, \ldots, x_n, \ldots)$ whose terms are such that for some p, $1 \leqslant p < \infty$, the series $\sum_{n=1}^{\infty} |\ x_n\ |^p$ converge. Let us introduce a metric in this set with the aid of the relationship

$$\rho\,(x,\ y)_p = \Big(\sum_{n=1}^{\infty} |\ x_n - y_n\ |^p\Big)^{1/p}, \tag{8}$$

where $y = (y_1, \ldots, y_n, \ldots)$ also belongs to X. The convergence of the series in (8) and also the triangle inequality

$$\rho\,(x,\ y)_p = \Big(\sum_{n=1}^{\infty} |\ (x_n - z_n) - (z_n - y_n)\ |^p\Big)^{1/p} \leqslant \rho\,(x,\ z)_p + \rho\,(z,\ y)_p$$

follow from Theorem 4 proved in the preceding section. The axioms of identity and symmetry for metric (8) are obvious. The space obtained from the set X by introducing in it metric (8) will be denoted by l^p. The convergence of the sequence $\{x^{(k)}\}$ to the element x means that

$$\rho\,(x^{(k)},\ x)_p = \Big(\sum_{n=1}^{\infty} |\ x_n^{(k)} - x_n\ |^p\Big)^{1/p} \to 0 \quad \text{as} \quad k \to \infty.$$

Hence there follow the relationships

(a) $\lim\limits_{k\to\infty} x_n^{(k)} = x_n$ for all $n = 1, 2, \ldots$ and

(b) $\left(\sum\limits_{n=N+1}^{\infty} |x_n^{(k)}|^p\right)^{1/p} < \varepsilon$ for all $N > N_0(\varepsilon)$ and all $k = 1, 2, \ldots$.

These relationships show the uniformity of the estimate of the remainder terms of the series of the sequence $\{x^{(k)}\}$ under consideration.

If $p = 2$, then the space l^2 is called the *Hilbert coordinate space*. It is a generalization of n-dimensional Euclidean space E_n for the case $n = \infty$.

7. Continuity of Operator. Let there be given two metric spaces X_{ρ_X} and Y_{ρ_Y}, and operator A defined on a certain set $M \subset X$ with values in Y, i.e. such that if $x \in M$, then $y = Ax \in Y$. Similarly to the definition of continuity of a function of a real variable, let us introduce the following definition.

The operator A is said to be *continuous at the point* $x_0 \in M$ if $\forall \varepsilon > 0\ \exists \delta = \delta(\varepsilon)$ such that $\rho_Y(Ax, Ax_0) < \varepsilon$ for all $x \in M$ satisfying the condition $\rho_X(x, x_0) < \delta$.

This definition implies that if the operator A is continuous at the point x_0 and the sequence $\{x_n\} \subset M$ is such that $\lim\limits_{n\to\infty} \rho_X(x_n, x_0) = 0$, then also $\lim\limits_{n\to\infty} \rho_Y(Ax_n, Ax_0) = 0$. Conversely, if for any sequence $\{x_n\} \subset M$ such that $\lim\limits_{n\to\infty} \rho_X(x_n, x_0) = 0$ we have $\lim\limits_{n\to\infty} \rho_Y(Ax_n, Ax_0) = 0$, then the operator A is continuous at the point x_0.

Example. Show that in the space E_n the extension operator

$$A = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}$$

is continuous at any point of the space.

Indeed, let for a given $\varepsilon > 0$ and a given point $x^{(0)}$ the sequence $\{x^{(k)}\}$ be chosen so that for $k \geqslant k_0(\varepsilon)$ the following inequality is fulfilled:

$$\rho_{E_n}(x^{(k)}, x^{(0)}) = \sqrt{\sum_{\nu=1}^{n} (x_\nu^{(k)} - x_\nu^{(0)})^2}$$

$$\leqslant \sqrt{n} \max_{1 \leqslant \nu \leqslant n} |x_\nu^{(k)} - x_\nu^{(0)}| < \frac{\varepsilon}{\max\limits_{1 \leqslant k \leqslant n} |\lambda_k|}.$$

Then for $k \geqslant k_0(\varepsilon)$ we have the relationship

$$\rho_{E_n}(Ax^{(k)}, Ax^{(0)}) = \sqrt{\sum_{\nu=1}^{n} \lambda_\nu^2 (x_\nu^{(k)} - x_\nu^{(0)})^2}$$

$$\leqslant \max_{1 \leqslant \nu \leqslant n} |\lambda_\nu| \sqrt{n} \max_{1 \leqslant \nu \leqslant n} |x_\nu^{(k)} - x_\nu^{(0)}| < \varepsilon.$$

Consequently, the extension operator A is continuous at every point $x^{(0)}$ of the space E_n.

Sec. 3.3. COMPLETENESS OF METRIC SPACES

1. Definition. The definition of the limit of a sequence in a metric space was given in the preceding section. Now it is natural to give the answer to the question concerning the necessary and sufficient conditions of the existence of the limit, the analogue of Cauchy's criterion in a metric space. It is clear that any sequence $\{x_n\}$ converging to the limit x^* satisfies the necessary condition of Cauchy's criterion: $\forall \varepsilon > 0 \; \exists N = N(\varepsilon)$ such that the inequality $\rho(x_n, x_m) < \varepsilon$ takes place for all $n, m > N(\varepsilon)$. But if the metric $\rho(r_1, r_2) = |r_1 - r_2|$ is introduced in the set of rational numbers $\{r\}$, then a metric space R is obtained in which not every sequence $\{r_n\}$ satisfying the sufficient condition of Cauchy's criterion will have a limit in this space R. This example shows that in an arbitrary metric space there is no analogue of Cauchy's criterion. Therefore a problem is set to separate those metric spaces in which Cauchy's criterion takes place. Let us introduce the following definition.

A sequence of elements $\{x_n\}$ of the metric space X_ρ is said to be a *convergent in itself sequence* (or *Cauchy's sequence*)

if $\forall \varepsilon > 0 \; \exists N = N(\varepsilon)$ such that

$$\rho(x_n, x_m) < \varepsilon \quad \forall n, m > N(\varepsilon). \tag{1}$$

It is obvious that any sequence $\{x_n\} \subset X_\rho$ converging to the limit x^*, $x^* \in X_\rho$, is a Cauchy sequence.

The metric space X_ρ is said to be *complete* if each Cauchy's sequence converges to the limit x^* which is an element of the same space X_ρ.

2. Theorem on Nested Balls. To get a certain characteristic of complete metric spaces, let us prove that in such spaces there takes place an analogue of the theorem on nested intervals. Let us first introduce the following notions.

The point $a \in X_\rho$ is called the *limit point of the set* $M \subset X_\rho$ if any neighbourhood $S(a, r)$ of the point a contains at least one point of the set $M \setminus a$, that is,

$$S(a, r) \cap (M \setminus a) \neq \theta \quad \text{for any} \quad r > 0.$$

The *closure of the set* M denoted by $\overline{M}$ is defined as the union of the set M and the set A of all its limit points, that is, $\overline{M} = M \cup A$.

Theorem 1. *In a complete metric space* X_ρ, *the sequence of nested closed balls*

$$\bar{S}_1(a_1, r_1) \supset \bar{S}_2(a_2, r_2) \supset \ldots \supset \bar{S}_n(a_n, r_n) \supset \ldots \tag{2}$$

with radii r_n *tending to zero has one, and only one, point* a *belonging to all the balls.*

First of all, let us show that the sequence of the centres of the balls, that is, the sequence $a_1, a_2, \ldots, a_n, \ldots$, converges in itself. Indeed, since for any $p = 1, 2, \ldots$ the following imbedding is valid:

$$\bar{S}_{n+p}(a_{n+p}, r_{n+p}) \subset \bar{S}_n(a_n, r_n),$$

the distance between the centres of the balls satisfies the condition $\rho(a_{n+1}, a_n) \leqslant r_n$. But $r_n \to 0$ as $n \to \infty$, and therefore $\rho(a_{n+p}, a_n) < \varepsilon$ if only $n > N(\varepsilon)$ and $p = 1, 2, \ldots$, and this means that the sequence $\{a_n\}$ converges in itself. By the hypothesis, X_ρ is a complete space, therefore there exists an element a, $a \in X_\rho$, such that $a = \lim\limits_{n \to \infty} a_n$.

Let $k \geqslant 1$ be a fixed integer. It follows from (2) that all the points a_k, a_{k+1}, . . ., a_{k+n}, . . . belong to the ball $\overline{S}_k(a_k, r_k)$, and therefore the limit point

$$a = \lim_{n \to \infty} a_{k+n} \in \overline{S}_k(a_k, r_k)$$

also belongs to this ball; the first part of the theorem has been proved.

Let us suppose now that there exists an element $b \in X_\rho$, $b \neq a$, such that $b \in \overline{S}_n(a_n, r_n)$ for all $n = 1, 2, \ldots$. Since $b \neq a$, we have $\rho(a, b) = \delta > 0$. On the other hand, applying the triangle inequality, we obtain the estimate

$$\rho(a, b) \leqslant \rho(a, a_n) + \rho(a_n, b) < 2r_n \to 0 \text{ as } n \to \infty,$$

from which it follows that $\delta > 0$ is impossible. Consequently, $\delta = 0$, and, by the identity axiom for a metric, we conclude that $b = a$.

The hypothesis of Theorem 1 is, in a way, a characteristic of complete metric spaces. Indeed, the following statement takes place.

Theorem 2. *If in a metric space X_ρ any sequence of nested closed balls with radii tending to zero has a non-empty intersection, then X_ρ is complete.*

Let $\{x_n\}$ be a *fundamental sequence* from X_ρ, that is,

$$\forall \varepsilon > 0 \; \exists N = N(\varepsilon) \text{ such that } \rho(x_n, x_{n+p}) < \varepsilon$$
$$\text{for all } n \geqslant N(\varepsilon) \text{ and } p = 1, 2, \ldots.$$

We choose the sequences $\varepsilon_k = 1/2^k$ and $N_k = N(1/2^k)$, for which

$$\rho(x_{N_k}, x_{N_{k+p}}) < 1/2^k \; \forall \, p = 1, 2, \ldots. \tag{3}$$

Consider the sequence of closed balls

$$\overline{S}_1(x_{N_1}, 1/2^0), \; \overline{S}_2(x_{N_2}, 1/2), \ldots, \; \overline{S}_k(x_{N_k}, 1/2^{k-1}), \ldots. \tag{4}$$

Taking into account that $x_{N_{k+1}} = x_{N_{k+p_0}}$ for a certain p_0, for any $x \in \overline{S}_{k+1}(x_{N_{k+1}}, 1/2^k)$ we have the estimate

$$\rho(x, x_{N_k}) \leqslant \rho(x, x_{N_{k+1}}) + \rho(x_{N_{k+1}}, x_{N_k}) < \frac{1}{2^k} + \frac{1}{2^k} = \frac{1}{2^{k-1}}.$$

This means that $x \in \overline{S}_k\,(x_{N_k},\ 1/2^{k-1})$, that is, the following imbedding holds:

$$\overline{S}_{k+1}\,(x_{N_{k+1}},\ 1/2^k) \subset \overline{S}_k\,(x_{N_k},\ 1/2^{k-1}).$$

The radii of the balls (4) tend to zero, and since, by the hypothesis of the theorem, their intersection is non-empty, there exists a point x_0 belonging to all balls. Let us show that x_0 is the limit of the sequence $\{x_n\}$. Indeed, we have the inequality

$$\rho\,(x_n,\ x_0) \leqslant \rho\,(x_n,\ x_{N_k}) + \rho\,(x_{N_k},\ x_0).$$

But $x_0 \in \overline{S}_k\,(x_{N_k},\ 1/2^{k-1})$ and the distance from x_{N_k} to x_0 satisfies the condition

$$\rho\,(x_{N_k},\ x_0) < \frac{1}{2^{k-1}},$$

therefore, choosing $n > N_k$ and using condition (3), we get the relationship

$$\rho\,(x_n,\ x_0) < \frac{1}{2^k} + \frac{1}{2^{k-1}} \to 0 \quad \text{as} \quad k \to \infty.$$

Consequently, x_0 is the limit of the sequence $\{x_n\}$, that is, X_ρ is a complete space.

Example 1°. Show that the n-dimensional Euclidean space E_n with the metric $\rho\,(x,\ y) = \sqrt{\sum\limits_{\nu=1}^{n} (x_\nu - y_\nu)^2}$ is a complete space.

Let $\{x^{(k)}\}$ be a fundamental sequence, i.e.

$$\rho\,(x^{(k)},\ x^{(m)}) < \varepsilon \quad \text{for all} \quad k,\ m > N\,(\varepsilon).$$

Hence it follows that for any $\nu = 1,\ 2,\ \ldots,\ n$ the following inequalities hold true:

$$|\,x_\nu^{(k)} - x_\nu^{(m)}\,| \leqslant \sqrt{\sum_{j=1}^{n} (x_j^{(k)} - x_j^{(m)})^2} < \varepsilon \quad \forall\, k,\ m > N\,(\varepsilon).$$

This means that the sequences of real numbers $\{x_\nu^{(k)}\}$, $\nu = 1,\ 2,\ \ldots,\ n$, are fundamental sequences and, by virtue

of Cauchy's criterion, there exist the limits

$$\tilde{x}_\nu = \lim_{k\to\infty} x_\nu^{(k)}, \quad \nu = 1, 2, \ldots, n. \tag{5}$$

Let us denote the sequence of these limit values by $\tilde{x}$. Then

$$\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) \in E_n$$

and we have the equality

$$\rho(\tilde{x}, x^{(k)}) = \sqrt{\sum_{j=1}^{n} (\tilde{x}_j - x_j^{(k)})^2}.$$

From relationships (5) we conclude that $\rho(\tilde{x}, x^{(k)}) \to 0$ as $k \to \infty$, i.e. E_n is a complete space.

Example 2°. Prove that the space $C[a, b]$ is complete.

Really, let $x_n(t) \subset C[a, b]$, $n = 1, 2, \ldots$, and let $\{x_n(t)\}$ be a fundamental sequence, that is, such that

$$\max_{a\leqslant t\leqslant b} |x_n(t) - x_m(t)| < \varepsilon \quad \text{if} \quad n, m > N(\varepsilon). \tag{6}$$

But condition (6) is the condition of a uniform convergence of the sequence of continuous functions $\{x_n(t)\}$, whose limit—the function $\tilde{x}(t)$—is also a continuous function: $\tilde{x}(t) \in C[a, b]$. Thus, any fundamental sequence $\{x_n(t)\}$ is convergent, and the space $C[a, b]$ is complete.

Example 3°. Prove that the space of convergent sequences a is complete.

Let $\{a^{(k)}\}$ be a fundamental sequence from c, i.e. such that

$$\rho(a^{(k)}, a^{(m)}) < \varepsilon \quad \text{for all} \quad k, m > N(\varepsilon).$$

Then from the definition of metric in c there follow the inequalities

$$\sup_n |a_n^{(k)} - a_n^{(m)}| < \varepsilon \quad \text{for all } k, m > N(\varepsilon),$$

and, therefore, also the inequalities

$$|a_n^{(k)} - a_n^{(m)}| < \varepsilon \quad \text{for} \quad k, m > N(\varepsilon) \quad \text{for all } n = 1, 2, \ldots.$$

From these estimates we deduce the existence of the limit with respect to each coordinate, that is, the existence of

the numbers

$$\alpha_n = \lim_{k\to\infty} a_n^{(k)}, \quad n = 1, 2, \ldots . \tag{7}$$

Thus, we have the sequence $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots)$. Let us prove that $\alpha \in c$. The following estimate is obvious:

$$|\alpha_n - \alpha_m| = |\alpha_n - a_n^{(k)} + a_n^{(k)} - a_m^{(k)} + a_m^{(k)} - \alpha_m|$$
$$\leqslant |\alpha_n - a_n^{(k)}| + |a_n^{(k)} - a_m^{(k)}| + |a_m^{(k)} - \alpha_m|.$$

From the fact that α_n is the limit of the sequence $\{a_n^{(k)}\}_{k=1}^{\infty}$ it follows that for $k > k_0(\varepsilon)$ the following inequalities hold:

$$|\alpha_n - a_n^{(k)}| < \varepsilon \quad \text{and} \quad |a_m^{(k)} - \alpha_m| < \varepsilon.$$

Further, since $a^{(k)} \in c$, for n, $m > N(\varepsilon)$ the following relationship is fulfilled:

$$|a_n^{(k)} - a_m^{(k)}| < \varepsilon.$$

Consequently, for n, $m > N(\varepsilon)$ the following estimate takes place:

$$|\alpha_n - \alpha_m| < 3\varepsilon$$

and, by Cauchy's criterion, $\alpha \in c$. Finally, from (7) it follows that

$$\rho(a^{(k)}, \alpha)_c = \sup_n |a_n^{(k)} - \alpha_n| < \varepsilon \quad \text{for} \quad k > k_0(\varepsilon),$$

that is,

$$\lim_{k\to\infty} a^{(k)} = \alpha \in c,$$

which proves the completeness of the space c.

The spaces $L^p[a, b]$ and l^p, $p \geqslant 1$, are also complete, but we are not going to prove this.

3. On Completion of Metric Spaces. The completeness of the real line plays an important role in mathematical analysis. In this case the set of real numbers is a supplement for the set of rational numbers such that the distances between elements in completion are preserved. Therefore, it is natural to set the problem of analogous completion of incomplete metric spaces. For a more precise formulation of this problem, let us introduce the following definitions.

The metric spaces X and Y are called *isometric* if a one-to-one correspondence can be established between them such that the distance between the preimages x_1 and x_2 in X is equal to the distance between the images y_1 and y_2 in Y, that is,

$$\rho_X(x_1, x_2) = \rho_Y(y_1, y_2).$$

The set M is said to be *dense in the set* G if its closure $\overline{M}$ contains G, i.e. $\overline{M} \supset G$.

The set M is *dense everywhere in the space* X if $\overline{M} = X$.

The following assertion is given without proof.

Let X_0 be an incomplete metric space. Then we have a complete metric space X such that it contains a subspace X_1 dense everywhere in X and isometric to the space X_0. The space X is called the supplement of the space X_0.

Example 4°. Show that the space P of given on $[-1, 1]$ algebraic polynomials $\{p(x) = \sum_{k=0}^{n} a_k x^k\}$ with the metric

$$\rho(p, q) = \max_{-1 \leqslant x \leqslant 1} |p(x) - q(x)|$$

is not complete.

Consider the sequence of polynomials

$$p_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}, \quad n = 0, 1, \ldots .$$

The limit of this sequence, the function e^x, does not belong to P.

From Weierstrass' theorem it follows that the space P is dense everywhere in the space of continuous functions $C[-1, 1]$, and this means that the space $C[-1, 1]$ can be regarded as the supplement of the space P.

Sec. 3.4. CONTRACTION MAPPING PRINCIPLE AND ITS APPLICATION

1. Contraction Operator Theorem. The method of successive approximations is frequently used in mathematical analysis, linear algebra, and in a number of other branches

of mathematics. It is widely applied in proving theorems on the existence of solutions of differential and integral equations and for obtaining approximate solutions of various equations. A generalization of this method for the case of operator equations in complete metric spaces was suggested by Stefan Banach. This generalization is called the *contraction mapping principle*. In what follows we shall prove this principle.

Theorem 1. *Let in a complete metric space X_ρ there be given an operator A which maps X into X so that $\forall\, x$ and $y \in X$*

$$\rho\,(Ax,\ Ay) \leqslant \alpha\rho\,(x,\ y), \tag{1}$$

where $0 < \alpha < 1$ and α is independent of x and y. Then there is one, and only one, point $x_0 \in X$ such that

$$Ax_0 = x_0. \tag{2}$$

In this case for the iterative process $x_{n+1} = Ax_n$, $n = 1, 2, \ldots$, taken beginning with the element $x \in X$, we have the following estimate for the nth approximation:

$$\rho\,(x_n,\ x_0) \leqslant \frac{\alpha^n}{1-\alpha}\,\rho\,(x,\ Ax). \tag{3}$$

The point $x_0 \in E$ satisfying condition (2) is called a *fixed point* of the operator A.

Let $x \in X$. Let us form the sequence

$$x_1 = Ax, \quad x_2 = Ax_1, \ \ldots, \ x_{n+1} = Ax_n, \ \ldots$$

and show that it converges in itself. Using condition (1), we have the inequalities

$$\rho\,(x_1,\ x_2) = \rho\,(Ax,\ Ax_1) \leqslant \alpha\rho\,(x,\ x_1) = \alpha\rho\,(x,\ Ax),$$

$$\rho\,(x_2,\ x_3) = \rho\,(Ax_1,\ Ax_2) \leqslant \alpha\rho\,(x_1,\ x_2) \leqslant \alpha^2\rho\,(x,\ Ax),$$

. . . . . .

$$\rho\,(x_n,\ x_{n+1}) \leqslant \alpha^n\rho\,(x,\ Ax),$$

. . . . . .

From these inequalities there follow the estimates

$$\rho\,(x_n,\ x_{n+p}) \leqslant \rho\,(x_n,\ x_{n+1}) + \rho\,(x_{n+1},\ x_{n+2}) + \ldots + \rho\,(x_{n+p-1},\ x_{n+p})$$

$$\leqslant (\alpha^n + \alpha^{n+1} + \ldots + \alpha^{n+p-1})\,\rho(x,\ Ax)$$
$$= \frac{\alpha^n - \alpha^{n+p}}{1-\alpha}\,\rho(x,\ Ax) \leqslant \frac{\alpha^n}{1-\alpha}\,\rho(x,\ Ax) \to 0$$
$$\text{as } n \to \infty, \quad (4)$$

where in the case $Ax = x$, in relationship (4) we have the equality sign. By the hypothesis, X is complete, therefore there exists an element $x_0 \in X$ such that $\lim\limits_{n\to\infty} x_n = x_0$.

Let us evaluate Ax_0. We have the inequalities

$$\rho(x_0,\ Ax_0) \leqslant \rho(x_0,\ x_n) + \rho(x_n,\ Ax_0)$$
$$= \rho(x_0,\ x_n) + \rho(Ax_{n-1},\ Ax_0)$$
$$\leqslant \rho(x_0,\ x_n) + \alpha\rho(x_{n-1},\ x_0) < \varepsilon$$

for all $n > n_0(\varepsilon)$. But the left-hand side, that is, $\rho(x_0, Ax_0)$ is independent of n, and therefore $\rho(x_0,\ Ax_0) = 0$, and from the metric property we conclude that $Ax_0 = x_0$.

Let us suppose now that there exists such an element $y_0 \in X$, $y_0 \neq x_0$, for which $Ay_0 = y_0$; then

$$\rho(x_0,\ y_0) = \rho(Ax_0,\ Ay_0) \leqslant \alpha\rho(x_0,\ y_0).$$

But since $\alpha < 1$, the last inequality occurs only for $\rho(x_0,\ y_0) = 0$, and hence there follows the equality $x_0 = y_0$, that is, the fixed point of the operator A is unique. Rewriting inequality (4) in the form

$$\rho(x_n,\ x_{n+p}) \leqslant \frac{\alpha^n}{1-\alpha}\,\rho(x,\ Ax)$$

and making p tend to infinity, we find the estimate

$$\rho(x_n,\ x_0) \leqslant \frac{\alpha^n}{1-\alpha}\,\rho(x,\ Ax),$$

and this is formula (3) for estimating the error. The estimate depends on the choice of the initial element $x \in X$.

2. Application of Contraction Mapping Principle to Solution of Equations. The contraction mapping principle is widely applied in constructing iterative processes for solving functional and differential equations. Let us consider several examples.

Example 1°. Find the solution of the equation $f(x) = x$ if $f(x)$ is a function differentiable at $x \in [a, b]$ which maps the interval $[a, b]$ into the interval $[a, b]$.

If $|f'(x)| \leqslant \alpha < 1$, then f is a contraction operator. This follows from the inequality

$$|f(x'') - f(x')| = |f'(\tilde{x})|\,|x'' - x'| \leqslant \alpha\,|x'' - x'|.$$

By Theorem 1, there exists a unique solution of the equation $f(x) = x$ which can be found using the iteration method. This iteration process is represented in Fig. 37. For an arbitrary element $\tilde{x}$ we find $f(\tilde{x})$, and from the point $(\tilde{x}, f(\tilde{x}))$ we draw a straight line $y = f(\tilde{x})$ to intersect the straight line $y = x$. Thus, we obtain the point $x_1 = f(\tilde{x})$. Further, we take $x_2 = f(x_1)$, $x_3 = f(x_2)$, etc. In the limit, we obtain a point x_0 such that $f(x_0) = x_0$.

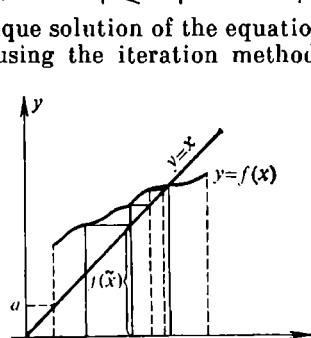


Fig. 37

Example 2°. Find the conditions under which the contraction mapping principle is applicable to solving the system of n linear algebraic equations

$$Ax = b,$$

where

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Consider the operator

$$Ux = (-A + I)\,x + b = Cx + b,$$

where I is a unit matrix. Then the solution of system (5) is a fixed point of the operator U which maps the n-dimensional vectors into n-dimensional vectors. Let us first consider the n-dimensional vectors in the space E_n with

the metric

$$\rho(x, y) = \sqrt{\sum_{\nu=1}^{n} (x_\nu - y_\nu)^2}.$$

Let us find the conditions under which U is a contraction operator. Applying Cauchy's inequality to the inner sum, we get the estimate

$$\rho(Ux', Ux'') = \sqrt{\sum_{k=1}^{n} \left[\sum_{j=1}^{n} c_{kj}(x'_j - x''_j)\right]^2}$$

$$\leqslant \sqrt{\sum_{k=1}^{n} \left[\sum_{j=1}^{n} c_{kj}^2 \sum_{j=1}^{n} (x'_j - x''_j)^2\right]}$$

$$= \rho(x', x'') \sqrt{\sum_{k=1}^{n} \sum_{j=1}^{n} c_{kj}^2},$$

from which it follows that if

$$\alpha = \sum_{k=1}^{n} \sum_{j=1}^{n} c_{kj}^2 < 1, \quad \text{where } c_{kj} = \begin{cases} -a_{kj} & \text{for } k \neq j, \\ 1 - a_{kk} & \text{for } j = k, \end{cases}$$

then system (5) has the unique solution which can be determined using the method of successive approximations.

Note that if the number of equations m in system (5) is less than the number of unknowns n ($a_{kj} = 0$ for $k = m + 1, \ldots, n$), then $c_{kj} = 1$ for $k = m + 1, \ldots, n$, and therefore $\alpha \geqslant 1$, and the contraction mapping principle is inapplicable here.

If we consider the set of n-dimensional vectors in the space E_n^* with the metric

$$\rho_1(x, y) = \max_{1 \leqslant k \leqslant n} |x_k - y_k|,$$

then for the images of the elements x' and x'' we obtain the relationship

$$\rho_1(Ux', Ux'') = \max_{1 \leqslant k \leqslant n} \left|\sum_{j=1}^{n} c_{kj}(x'_j - x''_j)\right|$$

$$\leqslant \max_{1 \leqslant k \leqslant n} \sum_{j=1}^{n} |c_{kj}| \max_{1 \leqslant j \leqslant n} |x'_j - x''_j|$$

$$= \rho_1(x', x'') \max_{1 \leqslant k \leqslant n} \sum_{j=1}^{n} |c_{kj}|,$$

whence it follows that if

$$\alpha_1 = \max_{1 \leqslant k \leqslant n} \sum_{j=1}^{n} |c_{kj}| < 1,$$

then system (5) can be solved by applying the contraction mapping principle.

This principle is also applicable in the case when a closed ball $\overline{S}(a, r)$ of a complete metric space X is mapped by the operator A into itself, that is, when $A\overline{S} \subset \overline{S}$.

Example 3°. Let the function $K(s, t, x(t))$ be continuous for $a \leqslant s \leqslant b$, $a \leqslant t \leqslant b$, and such that it satisfies a Lipschitz condition with respect to the argument x, that is, for any continuous on $[a, b]$ functions $x_1(t)$ and $x_2(t)$ the following inequality is fulfilled:

$$|K(s, t, x_1(t)) - K(s, t, x_2(t))| \leqslant M |x_1(t) - x_2(t)|, \quad (6)$$

where the constant M is independent of s and t. Show that for solving the integral equation

$$x(s) = \lambda \int_a^b K(s, t, x(t))\, dt \quad (7)$$

on the subset of functions $x(t)$ continuous on $[a, b]$ and satisfying the condition $|x(t)| \leqslant H$ for

$$|\lambda| < \frac{1}{b-a} \min\left(\frac{1}{M}, \frac{H}{\max |K(s, t, x)|}\right) \quad (8)$$

the contraction mapping principle is applicable.

Let $\overline{S}_H$ be a closed ball in $C[a, b]$, that is, the set of functions from $C[a, b]$ such that $|x(t)| \leqslant H$, and let

$$Ux = \lambda \int_a^b K(s, t, x(t))\, dt.$$

Then the following inequality holds true:

$$|Ux| = |\lambda| \left| \int_a^b K(s, t, x(t))\, dt \right| \leqslant |\lambda| \max |K(s, t, x(t))| (b - a).$$

By virtue of (8), we obtain the estimate $|Ux| < H$, from which we conclude that the operator U maps the set $\bar{S}_H$ into itself.

Further, using condition (6), we derive the relationship

$$|Ux - U\tilde{x}| \leqslant |\lambda| \int_a^b |K(s, t, x) - K(s, t, \tilde{x})| \, dt$$

$$\leqslant |\lambda| M \int_a^b |x(t) - \tilde{x}(t)| \, dt$$

$$\leqslant |\lambda| M \int_a^b \max_{a \leqslant t \leqslant b} |x(t) - \tilde{x}(t)| \, dt,$$

which we will write in the form

$$|Ux - U\tilde{x}| \leqslant |\lambda| M \rho_C(x, \tilde{x})(b - a) < \alpha \rho(x, \tilde{x}),$$

where, by virtue of condition (8), for the quantity α the following estimate is valid:

$$\alpha = |\lambda| M (b - a) < 1.$$

This means that the contraction mapping method is applicable for solving integral equation (7).

Sec. 3.5. COMPACT SETS

1. Definition. It is well known that of any bounded infinite set of points of the number line it is possible to separate at least one convergent sequence. The importance of this statement for a rigorous foundation of mathematical analysis was underlined by the Czechoslovakian mathematician Bernhard Bolzano at the beginning of the nineteenth century, and a more precise formulation of this statement was suggested by the outstanding German mathematician Karl Weierstrass. The idea of separation of a convergent sequence from sets in metric spaces was utilized by us, for instance, in the preceding section for proving the existence of a solution of a system of n linear equations and for proving the existence of a solution of an integral equation. Here, we are

concerned with the separation of those sets in a metric space for which there are analogues of the Bolzano-Weierstrass theorem. The condition of the boundedness of a set is no longer sufficient for this. Let us illustrate this by an example.

Example. Show that of the bounded in the space l^2 set of vectors

$$\mathbf{e}^{(1)} = (1, 0, \ldots, 0, \ldots), \quad \mathbf{e}^{(2)} = (0, 1, 0, \ldots, 0, \ldots), \quad \ldots, \quad \mathbf{e}^{(n)} = (\underbrace{0, \ldots, 0, 1}_{n}, 0, \ldots)$$

it is impossible to separate a convergent subsequence.

The set is bounded, since

$$\rho(\mathbf{e}^{(k)}, \theta) = \sqrt{\sum_{n=1}^{\infty} (e_n^{(k)} - 0)^2} = 1 \quad \text{for any } k = 1, 2, \ldots.$$

But for any $k \neq m$ we have the equalities

$$\rho(\mathbf{e}^{(k)}, \mathbf{e}^{(m)}) = \sqrt{\sum_{n=1}^{\infty} (e_n^{(k)} - e_n^{(m)})^2} = \sqrt{2},$$

and this means that of the infinite sequence $\{\mathbf{e}^{(k)}\}$ it is impossible to separate a convergent subsequence.

The sets of a metric space for which there are analogues of the Bolzano-Weierstrass theorem are called compact sets. Let us give their definition.

The set K of a metric space X is said to be *compact* in X or simply a *compact* if from any infinite sequence $\{x_n\} \subset K$ it is possible to separate a subsequence $\{x_{n_k}\}$ which converges to some limit $x_0 \in X$.

The set K of a metric space X is said to be *compact in itself* if from any infinite sequence $\{x_n\} \subset K$ it is possible to separate a subsequence $\{x_{n_k}\}$ converging to the element x_0 from the same set K.

A metric space X is said to be compact if from any infinite sequence $\{x_n\} \subset X$ it is possible to separate a subsequence $\{x_{n_k}\}$ converging to the element $x_0 \in X$.

As it was established in the example, the set of vectors $\{\mathbf{e}^{(k)} = (0, \ldots, 0, 1, 0, \ldots)\}$ lying in the closed unit ball $\bar{S}(\theta, 1)$ of the space l^2 is not compact. But a closed

unit ball in the n-dimensional Euclidean space E_n, that is, the set of vectors $x\ (x_1, \ldots, x_n)$ whose coordinates satisfy the condition $\sum_{k=1}^{n} x_k^2 \leqslant 1$, is compact in itself. This follows from the fact that the Bolzano-Weierstrass theorem is valid with respect to each coordinate x_ν.

Although the condition of boundedness of the set K is not sufficient for this set to be compact, nevertheless it is necessary. This fact is the consequence of the following theorem.

Theorem 1. *The compact set K of a metric space X_ρ can be loaded into a ball $S\ (a, r)$, $a \in X$, of a finite radius $r > 0$.*

Suppose the statement is false. This means that there exist a sequence $\{x_n\} \subset K$ and the point $a \in X$ such that $\rho\ (a, x_n) = r_n - 1$, where $r_n \to \infty$. We may regard that $r_{n+1} - 1 \geqslant r_n$, otherwise we would choose a subsequence of the sequence $\{r_n\}$. Let x_n and x_m be points of our sequence. Then the following relationships hold true:

$$\rho\ (a, x_n) = r_n - 1 \geqslant r_{n-1} \text{ and } \rho\ (a, x_m) = r_m - 1 \geqslant r_{m-1}.$$

Applying the triangle inequality, we have the estimate

$$\rho\ (a, x_n) \leqslant \rho\ (a, x_m) + \rho\ (x_m, x_n). \qquad (1)$$

If $n > m$, then $\rho\ (a, x_n) = r_n - 1 \geqslant r_{n+1} \geqslant r_m$, therefore from estimate (1) we obtain the inequality

$$r_m \leqslant r_m - 1 + \rho\ (x_n, x_m), \quad \text{i.e.} \quad \rho\ (x_n, x_m) \geqslant 1.$$

Hence it follows that it is impossible to separate a convergent subsequence from the sequence $\{x_n\}$. Thus, we have arrived at a contradiction to our supposition that K is compact. Consequently, the set K can be loaded into the ball $S\ (a, r)$ of radius $r > 0$.

2. Hausdorff's Theorem. The condition of boundedness of the set K of the metric space X_ρ is a necessary but insufficient condition of compactness, therefore it is natural to pose the problem of obtaining the criterion of compactness of the set K. The most general such criterion is Hausdorff's criterion for formulation of which the following definition should be introduced.

The set A, $A \subset K \subset X_\rho$, is referred to as the *ε-net*, $\varepsilon > 0$, for the set K if for any element $x \in K$ there is an element $y \in A$ such that $\rho(x, y) < \varepsilon$.

Theorem 2 (Hausdorff's Theorem). *For a set K to be compact in a metric space X it is necessary, and if X is complete, then also sufficient that for any $\varepsilon > 0$ in the set K there is a finite ε-net for K.*

Necessity. Let K be compact and for some $\varepsilon > 0$ there is no finite ε-net. We choose $x_1 \in K$, then $\exists\, x_2 \in K$ such that $\rho(x_1, x_2) \geqslant \varepsilon$ (otherwise x_1 would be a finite ε-net consisting of one element). Let there be determined the elements $x_1, x_2, \ldots, x_n$ such that $\rho(x_i, x_k) \geqslant \varepsilon$ for all $i \neq k$, $i, k = 1, 2, \ldots, n$. Since a finite number of points is not an ε-net, $\exists\, x_{n+1} \in K$ such that $\rho(x_i, x_{n+1}) \geqslant \varepsilon$ for all $i = 1, 2, \ldots, n$. Continuing successively the choice of the elements $x_{n+2}, x_{n+3}, \ldots$, we obtain an infinite sequence $x_1, x_2, \ldots, x_n, \ldots$ such that $\rho(x_i, x_k) \geqslant \varepsilon$ for all $i \neq k$, $i, k = 1, 2, \ldots$. But this means that from this sequence $\{x_n\}$ it is impossible to separate a convergent subsequence, which contradicts our supposition that K is compact. Thus, there exists a finite ε-net in K.

Sufficiency. Let X be a complete space, and let there exist a finite ε-net for any $\varepsilon > 0$ in K. Let us choose a sequence $\{\varepsilon_n\}$, $\lim\limits_{n\to\infty} \varepsilon_n = 0$, and for each ε_n construct a finite ε_n-net:

$$\begin{aligned}
&\text{for } \varepsilon_1 \quad \{x_1^{(1)}, x_2^{(1)}, \ldots, x_{k_1}^{(1)}\} = A_1,\\
&\text{for } \varepsilon_2 \quad \{x_1^{(2)}, x_2^{(2)}, \ldots, x_{k_2}^{(2)}\} = A_2,\\
&\qquad \ldots\ldots\ldots\ldots\ldots\ldots\\
&\text{for } \varepsilon_n \quad \{x_1^{(n)}, x_2^{(n)}, \ldots, x_{k_n}^{(n)}\} = A_n,\\
&\qquad \ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{2}$$

Let B be any infinite subset from K. About each of the points of the set $A_1 = \{x_\nu^{(1)}\}$ we describe closed balls of radius ε_1. Every point from B will be enclosed in one of these balls, therefore let us choose that ball $\overline{S}(x_{\nu_1}^{(1)}, \varepsilon_1)$ which contains the infinite part from B. We denote this part by B_1 and describe closed balls about each of the points belonging to the set $A_2 = \{x_\nu^{(2)}\}$ of radius ε_2. Let us choose the ball $\overline{S}(x_{\nu_2}^{(2)}, \varepsilon_2)$, which contains the infinite subset

$B_2 \subset B_1$. Continuing this process, we find the sequence of the closed balls

$$\overline{S}(x_{\nu_1}^{(1)}, \varepsilon_1), \overline{S}(x_{\nu_2}^{(2)}, \varepsilon_2), \ldots, \overline{S}(x_{\nu_n}^{(n)}, \varepsilon_n), \ldots$$

and the infinite sequence of infinite sets:

$$B \supset B_1 \supset B_2 \supset \ldots \supset B_n \supset \ldots .$$

Choosing the points $\eta_i \in B_i$, $\eta_i \neq \eta_j$, we obtain a sequence of points $\eta_1, \eta_2, \ldots, \eta_n, \ldots$, which converges in itself, since for $n > m$ the points η_n and η_m belong to the ball $\overline{S}(x_{\nu_m}^{(m)}, \varepsilon_m)$, and therefore $\rho(\eta_n, \eta_m) < \varepsilon_m$. The completeness of X implies that $\exists\ \eta_0 = \lim\limits_{n\to\infty} \eta_n$, $\eta_0 \in X$. Thus, in an infinite set B we have constructed the sequence $\{\eta_n\}$ converging to $\eta_0 \in X$, i.e. the set K is compact in X.

3. Arzela's Theorem. Let us establish the criterion of compactness of sets in the space $C[a, b]$. But first we are going to introduce the following definitions.

The family of functions $\Phi = \{x(t)\} \subset C[a, b]$ is called *equibounded* if there is such $M > 0$ that for all functions of this family Φ the inequality $|x(t)| \leqslant M$ holds true.

The family of functions $\Phi \subset C[a, b]$ is said to be *equicontinuous* if $\forall\ \varepsilon > 0\ \exists\ \delta = \delta(\varepsilon) > 0$ such that $\forall\ t_1, t_2 \in [a, b]$ for which $|t_1 - t_2| < \delta$, then for any function $x(t) \in \Phi$ the following inequality is fulfilled:

$$|x(t_1) - x(t_2)| < \varepsilon.$$

Theorem 3 (Arzela's Theorem). *For a set $K \subset C[a, b]$ to be compact in $C[a, b]$, it is necessary and sufficient that the set K be equibounded and equicontinuous.*

Necessity. Let the set K be compact. Then, by virtue of Theorem 1, in the space $C[a, b]$ there exists a ball $S(0, R)$ such that $K \subset S(0, R)$. This means that for any function $x(t) \in K$ we have the inequalities $-R \leqslant x(t) \leqslant R$, i.e. $|x(t)| \leqslant R$. Consequently, the set K is equibounded.

Further, let $\varepsilon > 0$ be arbitrary. We choose in K a finite ε-net: $y_1(t), y_2(t), \ldots, y_m(t)$. This finite set of functions is equicontinuous, and therefore there exists $\delta = \delta(\varepsilon) > 0$ such that for all $i = 1, 2, \ldots, m$ and for all $t_1, t_2 \in [a, b]$ satisfying the inequality $|t_1 - t_2| < \delta$ the follow-

ing inequalities hold:

$$| y_i(t_1) - y_i(t_2) | < \varepsilon, \quad i = 1, 2, \ldots, m.$$

If now $x(t)$ is an arbitrary function from K, then there is i_0 such that $\rho(x, y_{i_0}) < \varepsilon$. Thus,

$$| x(t_1) - x(t_2) | \leqslant | x(t_1) - y_{i_0}(t_1) | + | y_{i_0}(t_1) - y_{i_0}(t_2) | + | y_{i_0}(t_2) - x(t_2) | < 3\varepsilon$$

and equicontinuity has been proved.

Sufficiency. Without loss of generality, we may regard that $[a, b] \equiv [0, 1]$. Equicontinuity implies that $\forall \varepsilon > 0 \; \exists \delta = \delta(\varepsilon) > 0$ such that $| x(t_1) - x(t_2) | < \varepsilon$ if only $| t_1 - t_2 | < \delta$, $t_1, t_2 \in [0, 1]$, $x(t) \in K$. Let n be such that $1/n \leqslant \delta$. Let us divide the interval $[0, 1]$ into equal parts:

$$\Delta_k = \left[\frac{k}{n}, \frac{k+1}{n}\right], \quad k = 0, 1, \ldots, n-1,$$

then

$$| x(t_1) - x(t_2) | < \varepsilon \quad \text{for} \quad t_1, t_2 \in \Delta_k. \tag{3}$$

Let us define the function $x_n(t)$ in the following way:

$$x_n(t) = \begin{cases} x(k/n) & \text{for } t = k/n, \; k = 0, 1, \ldots, n-1, \\ \text{linear on } \Delta_k, \end{cases}$$

that is, $x_n(t)$ is a broken line inscribed in $x(t)$.

If $x(k/n) \leqslant x((k+1)/n)$, then from the linearity of the function $x_n(t)$ there follow the estimates

$$x\left(\frac{k}{n}\right) \leqslant x_n(t) \leqslant x\left(\frac{k+1}{n}\right) \quad \text{for} \quad t \in \Delta_k.$$

Subtracting these inequalities from $x(t)$ and using condition (3), we come to the relationship

$$-\varepsilon < x(t) - x\left(\frac{k+1}{n}\right) \leqslant x(t) - x_n(t) \leqslant x(t) - x\left(\frac{k}{n}\right) < \varepsilon.$$

A similar inequality also exists in the case $x(k/n) > x((k+1)/n)$. Thus, the following estimate is valid:

$$| x(t) - x_n(t) | \leqslant \varepsilon \quad \text{for all} \quad t \in [0, 1], \tag{4}$$

whence there follows the relationship

$$|x_n(t)| \leqslant |x_n(t) - x(t)| + |x(t)| < \varepsilon + M = M_1,$$

and this means that the set $N = \{x_n(t)\}$ is bounded.

Let us associate each function $x_n(t) \in N$ with a point of the $(n+1)$-dimensional space E_{n+1}, that is, with the point z_n:

$$x_n(t) \to z_n, \quad z_n = (x_n(0), x_n(1/n), \ldots, x_n(1)) \in E_{n+1}.$$

The coordinates z_n coincide with the values $x_n(t)$ at the points $t = k/n$. This correspondence is one-to-one and continuous. Indeed, if in the metric of $C[0, 1]$ for the points $x_n^{(1)}(t)$ and $x_n^{(2)}(t)$ we have the inequality $\rho_C(x_n^{(1)}, x_n^{(2)}) < \varepsilon$, then for their images $z_n^{(1)}$ and $z_n^{(2)}$ in the space E_{n+1} we obtain the estimate

$$\rho_{E_{n+1}}(z_n^{(1)}, z_n^{(2)}) = \sqrt{\sum_{\nu=0}^{n} \left[x_n^{(1)}\left(\frac{\nu}{n}\right) - x_n^{(2)}\left(\frac{\nu}{n}\right)\right]^2}$$

$$\leqslant \sqrt{n}\rho_C(x_n^{(1)}, x_n^{(2)}) = \varepsilon\sqrt{n} = \varepsilon_1. \qquad (5)$$

Conversely: if the images are close, that is,

$$\rho_{E_{n+1}}(z_n^{(1)}, z_n^{(2)}) = \sqrt{\sum_{\nu=0}^{n} \left[x_n^{(1)}\left(\frac{\nu}{n}\right) - x_n^{(2)}\left(\frac{\nu}{n}\right)\right]^2} < \varepsilon,$$

then for any $\nu = 0, 1, \ldots, n$ we have the inequalities

$$\left|x_n^{(1)}\left(\frac{\nu}{n}\right) - x_n^{(2)}\left(\frac{\nu}{n}\right)\right| < \varepsilon. \qquad (6)$$

But then for $t \in \Delta_\nu$ from the relationships (3), (4), and (6) there follows the estimate

$$|x_n^{(1)}(t) - x_n^{(2)}(t)|$$

$$\leqslant \left|x_n^{(1)}(t) - x_n^{(1)}\left(\frac{\nu}{n}\right)\right| + \left|x_n^{(1)}\left(\frac{\nu}{n}\right) - x_n^{(2)}\left(\frac{\nu}{n}\right)\right|$$

$$+ \left|x_n^{(2)}\left(\frac{\nu}{n}\right) - x_n^{(2)}(t)\right| < 3\varepsilon,$$

which means that

$$\rho_C(x_n^{(1)}, x_n^{(2)}) < 3\varepsilon. \qquad (7)$$

Thus, the correspondence $x_n(t) \leftrightarrow z_n$ is continuous, that is, the proximity of the points $x_n^{(1)}(t)$ and $x_n^{(2)}(t)$ in the metric of $C[0, 1]$ implies the proximity of their images $z_n^{(1)}$ and $z_n^{(2)}$ in the metric of E_{n+1}, and conversely.

The set $N = \{x_n(t)\}$ is bounded in the space $C[0, 1]$, therefore the set of images $\widetilde{N} = \{z_n\}$ is bounded in the space E_{n+1} and, by virtue of the Bolzano-Weierstrass theorem, is compact in E_{n+1}.

According to Hausdorff's theorem, for a given $\varepsilon > 0$ in the set $\widetilde{N}$ there exists a finite ε-net $\{z_{n,1}, z_{n,2}, \ldots, z_{n,m(\varepsilon)}\}$. Let $x_{n,1}(t), x_{n,2}(t), \ldots, x_{n,m(\varepsilon)}(t)$ be their preimages in N. From the continuity of the correspondence $N \leftrightarrow \widetilde{N}$, that is, from relationships (5) and (7), it follows that $x_{n,1}(t), \ldots, x_{n,m(\varepsilon)}(t)$ is a finite 3ε-net for the set N.

Let now $x(t)$ be an arbitrary element from K, and let $\varepsilon > 0$ be arbitrary. Then from condition (4) we conclude that there exists a function $x_n^{(\varepsilon)}(t) \in N$ such that $|x(t) - x_n^{(\varepsilon)}(t)| < \varepsilon$. But $\{x_{n,\nu}(t)\}_{\nu=1}^{m(\varepsilon)}$ is a finite ε-net for N, and this means that there is a function $x_{n,\mu}(t)$ such that $|x_{n,\mu}(t) - x_n^{(\varepsilon)}(t)| < \varepsilon$. Thus, the estimate $|x(t) - x_{n,\mu}(t)| < 2\varepsilon$ holds which means that the set $\{x_{n,\nu}(t)\}_{\nu=1}^{m(\varepsilon)}$ is a 2ε-net for the set K. By Hausdorff's theorem, we conclude that the set K is compact in $C[0, 1]$.

4. Separability and Compactness. When performing computations in the set of real numbers, irrational numbers are frequently replaced with any degree of accuracy by rational numbers whose number is countable. There arises the question of separating such dense countable sets in other metric spaces as well. The generalization of the property of density of the set of rational numbers in the set of real numbers in metric spaces is separability.

The space X_ρ is called *separable* if there exists in it a countable everywhere dense set, that is, there exists a sequence $\{x_n\} \subset X$ such that for each $x \in X$ and any $\varepsilon > 0$ there is an element x_{n_1}, $x_{n_1} \in \{x_n\}$, for which $\rho(x, x_{n_1}) < \varepsilon$.

A separable space is the space E_n in which a countable everywhere dense set is the set of points with rational coordinates.

Let us note, without proof, that the spaces $C[a, b]$, $L^p[a, b]$, and l^p, $1 \leqslant p < \infty$, are also separable.

Compactness is one of the criteria of separability.

Theorem 4. *A compact space X is separable.*

We choose a sequence $\{\varepsilon_n\}$, $\varepsilon_n > 0$, $\lim\limits_{n\to\infty} \varepsilon_n = 0$ and for each ε_n construct an ε_n-net (see (2)). Then the union $A = \bigcup\limits_{n=1}^{\infty} A_n$ is countable and everywhere dense in X. Indeed for any $x \in X$ and an arbitrary $\varepsilon > 0$ we choose a number n_1 so that $\varepsilon_{n_1} \leqslant \varepsilon$. Now, we choose an element $x_\mu^{(n_1)} \in A_{n_1} \subset A$ such that the relationship $\rho(x, x_\mu^{(n_1)}) < \varepsilon_{n_1} \leqslant \varepsilon$ is fulfilled, and this just shows that the set A is dense in X.

CHAPTER 4

Completely Continuous Operators in Normed Linear Spaces

Sec. 4.1. NORMED LINEAR SPACES

1. Definition. The set E of elements is called a *normed linear space* provided that the following conditions are fulfilled:

(1) The set E is an *Abelian group* with respect to the operation of addition, that is, if $x \in E$, $y \in E$, then also $x + y \in E$, where:

(a) $x + y = y + x$ (commutativity);

(b) $x + (y + z) = (x + y) + z$ (associativity);

(c) there exists an element θ such that $x + \theta = x$ for all $x \in E$;

(d) for each $x \in E$ there is an element $(-x) \in E$ such that $x + (-x) = \theta$.

(2) The multiplication of the elements of the set E by the numbers from the real (or complex) field D (or K) satisfying the following conditions is defined:

(e) $\lambda(\mu x) = (\lambda\mu)x$ (associativity of multiplication);

(f) $\left.\begin{array}{l}\lambda(x + y) = \lambda x + \lambda y \\ (\lambda + \mu)x = \lambda x + \mu x\end{array}\right\}$ (distribution laws);

(g) $1 \cdot x = x$.

(3) Each element x of the set E is associated with a real number $\| x \|$ called the *norm* of this element, and the norm satisfies the following axioms:

(h) $\| x \| \geqslant 0$, and $\| x \| = 0$ only for $x = \theta$;

(i) $\| x + y \| \leqslant \| x \| + \| y \|$;

(j) $\| \lambda x \| = | \lambda | \, \| x \|$.

In a normed linear space, it is possible to introduce the metric $\rho(x, y) = \| x - y \|$, the fulfilment of whose axioms is obvious. The convergence of the elements of $\{x_n\}$ with

respect to this metric, namely:

$$\rho(x_n, x) = \| x_n - x \| \to 0 \quad \text{as} \quad n \to \infty,$$

is called the *convergence in norm.*

If a normed linear space is complete in the sense of convergence in norm, then it is called the *space of type B*, or a *Banach space.*

Here are some examples of Banach spaces:

space E_n with norm $\| x \| = \left(\sum_{\nu=1}^{n} x_\nu^2 \right)^{1/2}$;

space $C[a, b]$ with norm $\| x \| = \max_{a \leqslant t \leqslant b} | x(t) |$;

spaces $L^p[a, b]$ and l^p, $p \geqslant 1$, with respective norms

$$\| x \|_{L^p} = \left(\int_a^b | x(t) |^p \, dt \right)^{1/p} \text{ and } \| x \|_{l^p} = \left(\sum_{\nu=1}^{\infty} | x_\nu |^p \right)^{1/p}.$$

2. Simplest Properties. Let us establish certain simplest properties of normed linear spaces.

Property 1. *If* $x_n \to x$, $y_n \to y$, *then* $x_n + y_n \to x + y$.

Indeed, from condition (i) we derive the relationships

$$\| (x + y) - (x_n + y_n) \| \leqslant \| x - x_n \| + \| y - y_n \| \to 0 \quad \text{as} \quad n \to \infty.$$

Property 2. *If* $x_n \to x$, $\lambda_n \to \lambda$, *then* $\lambda_n x_n \to \lambda x$.

Indeed, taking into consideration conditions (i) and (j), we obtain

$$\| \lambda_n x_n - \lambda x \| = \| \lambda_n (x_n - x) + (\lambda_n - \lambda) x \|$$
$$\leqslant | \lambda_n | \, \| x_n - x \| + | \lambda_n - \lambda | \, \| x \| \to 0 \quad \text{as} \quad n \to \infty.$$

Property 3. *If* $x_n \to x$, *i.e.* $\| x_n - x \| \to 0$, *then* $\| x_n \| \to \| x \|$. *In other words, the convergence in norm of a space implies the convergence of norms.*

Substituting the element x in the form $x = y + (x - y)$, we get the estimate

$$\| x \| \cdot \| y + (x - y) \| \leqslant \| y \| + \| x - y \|, \quad \text{or}$$
$$\| x \| - \| y \| \leqslant \| x - y \|.$$

Analogously, we find that $\| y \| - \| x \| \leqslant \| x - y \|$; thus,

$$| \| x \| - \| y \| | \leqslant \| x - y \|.$$

Using the inequality obtained, we derive the relationship

$$| \| x_n \| - \| x \| | \leqslant \| x_n - x \|.$$

But the right-hand side, by the hypothesis, tends to zero, and therefore $\| x_n \| \to \| x \|$.

Normed linear spaces are metric spaces, therefore they have all the properties of metric spaces.

3. Linear Manifolds. Convex Sets. The set of elements $x_1, x_2, \ldots, x_n$ from a normed linear space E is said to be *linearly independent* if it follows from the equality $\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n = \theta$ that $\lambda_1 = \lambda_2 = \ldots = \lambda_n = 0$.

A non-empty set L of elements of the linear space E is called a *linear manifold* if, together with the elements $x_1, x_2, \ldots, x_n$, it also contains every linear combination $\sum_{\nu=1}^{n} \alpha_\nu x_\nu$ of these elements. The simplest linear manifold of the normed linear space E is a straight line defined by the element x, that is, the set of elements of the form $y = tx$, $x \in E$, $-\infty < t < \infty$.

Any linear manifold necessarily contains the element θ. Indeed, the manifold L is non-empty, that is, $L \ni x$, and therefore we have $(-x) \in L$ $(\alpha = -1)$ and $x + (-x) = \theta \in L$ $(\alpha_1 = \alpha_2 = 1)$.

The number of linearly independent elements $x_1, x_2, \ldots, x_n$ generating a given manifold is called the *dimension* of this manifold.

If for any n in the normed linear space E there exists n linearly independent elements, then E is said to be *infinite dimensional*.

The set of elements of the form

$$y = tx_1 + (1 - t) x_2, \quad 0 \leqslant t \leqslant 1,$$

defined by the elements x_1 and $x_2 \in E$ is called the *interval*.

The set M is termed *convex* if each line segment joining two arbitrary points of the set M belongs entirely to the

set M, that is, from the condition $x_1,\ x_2 \in M$ it follows that $y = tx_1 + (1 - t)\, x_2 \in M$ for any $t \in [0,\ 1]$.

The intersection of convex sets is again a convex set. Consequently, for every set $A \subset E$ there exists a least convex set in E containing the set A. It is called the *convex hull of the set* A. The convex hull of the set A consists of all finite sums of the form $\sum_i \alpha_i x_i$, where x_i are arbitrary elements from A, $0 \leqslant \alpha_i \leqslant 1$, $\sum_i \alpha_i = 1$.

Theorem. *In a normed linear space, any closed bal* $\| x - a \| \leqslant r$ *is a convex set.*

Indeed, let

$$\| x_1 - a \| \leqslant r, \quad \| x_2 - a \| \leqslant r, \quad \text{and}$$
$$y = tx_1 + (1 - t)\, x_2, \quad 0 \leqslant t \leqslant 1.$$

Then, we have the relationships

$$\begin{aligned} \| y - a \| &= \| tx_1 + (1 - t)\, x_2 - a \| \\ &= \| tx_1 + (1 - t)\, x_2 - ta - (1 - t)\, a \| \\ &= \| t\,(x_1 - a) + (1 - t)\,(x_2 - a) \| \\ &\leqslant t \| x_1 - a \| + (1 - t) \| x_2 - a \| \\ &\leqslant tr + (1 - t)\, r = r \end{aligned}$$

that is, $\| y - a \| \leqslant r$, and this means that y belongs to the same ball.

Sec. 4.2. CONTINUOUS AND COMPLETELY CONTINUOUS OPERATORS

1. Definitions. Let A be an operator defined on the space E_x with values in the space E_y. The operator A is said to be *additive* if for any $x_1,\ x_2 \in E_x$ the relationship $A\,(x_1 + x_2) = Ax_1 + Ax_2$ is fulfilled.

The operator A is called *continuous at the point* $x_0 \in E_x$ if from the condition $x_n \to x_0$, or $\| x_n - x_0 \| \to 0$ as $n \to \infty$, it follows that $Ax_n \to Ax_0$ ($\| Ax_n - Ax_0 \| \to 0$ as $n \to \infty$).

It turns out that the requirement of the continuity of an additive operator at a point is a strong restriction. This is confirmed by the following theorem.

Theorem 1. *If an additive operator* A: $(E_x \to E_y)$ *is continuous at the point* $x_0 \in E_x$, *then it is also continuous on the whole* E_x.

Let x be an arbitrary point of the space E_x. We choose a sequence of points $x_n \in E_x$ so that $x_n \to x$. Then we have $(x_n - x + x_0) \to x_0$, and the continuity of the operator A at the point x_0 implies the following relationship:

$$\lim_{n\to\infty} A(x_n - x + x_0) = Ax_0. \tag{1}$$

From the additivity of the operator A we derive the equality

$$A(x_n - x + x_0) = Ax_n - Ax + Ax_0.$$

Taking into consideration this equality, from (1) we get the expression $\lim_{n\to\infty} Ax_n - Ax + Ax_0 = Ax_0$, wherefrom it follows that $\lim_{n\to\infty} Ax_n = Ax$.

In what follows, we shall be mainly concerned with completely continuous operators.

The operator A continuous on the set M and attaining values in the same set M is referred to as *completely continuous* if it maps any bounded set $M_1 \subset M$ into a compact set, that is, the set $A(M_1)$ is a compact set.

2. Continuity of the Limit of a Uniformly Convergent Sequence of Completely Continuous Operators. Completely continuous operators possess a number of properties used in applications. One of such properties is established by the following theorem.

Theorem 2. *If the sequence* $\{A_n\}$ *of operators that are completely continuous on a bounded set* M *converges uniformly on* M *to the operator* A_0, *i.e.* $\forall \varepsilon > 0\ \exists N = N(\varepsilon)$ *such that* $\| A_n x - A_0 x \| < \varepsilon$ *for all* $x \in M$, *if only* $n > N(\varepsilon)$, *then* A_0 *is completely continuous on* M.

It is necessary to prove that the operator A_0 is continuous on M and that $A_0(M) = N$ is a compact set. Let $\{x_m\} \subset M$, $x_m \to x_0 \in M$; then we can write the inequality

$$\| A_0 x_m - A_0 x_0 \| \leqslant \| A_0 x_m - A_n x_m \| + \| A_n x_m - A_n x_0 \| + \| A_n x_0 - A_0 x_0 \|.$$

The first and third terms on the right are less than $\varepsilon/3$ for $n > N(\varepsilon)$, by virtue of the uniform convergence of the sequence $\{A_n\}$, while the second term is less than $\varepsilon/3$ for $m > m_0(\varepsilon)$, since the operators A_n are continuous. Thus, for $m > m_0(\varepsilon)$ we have the estimate

$$\| A_0 x_m - A_0 x_0 \| < \varepsilon,$$

which means that the operator A_0 is continuous on M.

Now, let us prove that the set $N = A_0(M)$ is compact. By the hypothesis, the set M is bounded. Let $\varepsilon > 0$. We choose the number $n_0 = n_0(\varepsilon)$ so that to fulfill the inequality

$$\| A_{n_0} x - A_0 x \| < \varepsilon \ \forall\, x \in M, \tag{2}$$

which is possible by virtue of the uniform convergence of the sequence $\{A_n\}$. Let us set $N_0 = A_{n_0}(M)$ and show that a finite ε-net in N_0 is a finite ε-net in N.

Indeed, let $y \in N$ and $x \in M$ be one of the preimages of the element y, that is, $y = A_0 x$. Setting $y_0 = A_{n_0} x$, $y_0 \in N$, from condition (2) we derive the inequality

$$\| y - y_0 \| = \| A_0 x - A_{n_0} x \| < \varepsilon. \tag{3}$$

As the range of values of the operator A_{n_0}, the set N_0 is compact, therefore from inequality (3) we conclude that a finite ε-net in N_0 will be also a finite ε-net for the set N. Thus, the set N is compact.

3. Complete Continuity of Fredholm's Operator. As an example of application of Theorem 2, let us consider Fredholm's operator and prove that it is completely continuous.

Let the function $K(t, s)$ be quadratically summable in the square $a \leqslant t, s \leqslant b$, that is,

$$B^2 = \int_a^b \int_a^b |K(t, s)|^2 \, dt \, ds < \infty. \tag{4}$$

In the space $L^2[a, b]$ consider the operator T $T\varphi$ representable in the form

$$T\varphi = \int_a^b K(t, s) \varphi(s) \, ds, \quad \varphi \in L^2[a, b]. \tag{5}$$

The operator $T\varphi$, determined by this formula, will be called *Fredholm's operator* with the kernel $K(t, s)$.

First of all, let us show that $T\varphi \in L^2[a, b]$, i.e. that Fredholm's operator maps the space $L^2[a, b]$ into itself.

It follows from condition (4) that the integrals

$$\int_a^b |K(t, s)|^2 dt \quad \text{and} \quad \int_a^b |K(t, s)|^2 ds$$

are finite: the first almost for all s, and the second almost for all t. The following inequality is obvious:

$$|K(t, s)\varphi(s)| \leqslant \frac{1}{2}|K(t, s)|^2 + \frac{1}{2}|\varphi(s)|^2.$$

In it, the functions, on the right, are integrable with respect to s on the interval $[a, b]$; therefore expression (5) is defined almost for all $t \in [a, b]$. Besides, applying the Cauchy-Buniakowski inequality (see Sec. 3.1), we have the following estimate:

$$|T\varphi|^2 \leqslant \int_a^b |K(t, s)|^2 ds \int_a^b |\varphi(s)|^2 ds$$

$$= \|\varphi\|^2 \int_a^b |K(t, s)|^2 ds. \qquad (6)$$

Using condition (4), we make sure from inequality (6) that $|T\varphi|^2$ is integrable, and this means that $T\varphi \in L^2[a, b]$. On having integrated inequality (6), we make sure that $\|T\varphi\| \leqslant B\|\varphi\|$.

The quantity $\sup_{\|\varphi\| \leqslant 1} \|T\varphi\|$ is called the *norm of the operator* T and is denoted by $\|T\|_1$; it is clear that $\|T\|_1 \leqslant B$.

Theorem 3. *Fredholm's operator* (5) *is completely continuous.*

Let $\{\psi_n(t)\}_{n=1}^{\infty}$ be a complete orthonormal in the space $L^2[a, b]$ system of functions. Then the system of products $\{\psi_n(t)\psi_m(s)\}$ is complete in the space $L^2([a, b] \times [a, b])$, and therefore the kernel of Fredholm's operator $K(t, s)$ can be represented in the form of the series

$$\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_{nm}\psi_n(t)\psi_m(s) \qquad (7)$$

which is convergent in norm of the space $L^2([a, b] \times [a, b])$ to the function $K(t, s)$. This means that if

$$K_N(t, s) = \sum_{n, m=1}^{N} a_{nm} \psi_n(t) \psi_m(s), \quad N = 1, 2, \ldots,$$

are partial sums of series (7), then

$$\int_a^b \int_a^b |K(t, s) - K_N(t, s)|^2 \, dt \, ds \to 0 \quad \text{for } N \to \infty. \tag{8}$$

Consider the operator T_N given in the form

$$T_N \varphi = \int_a^b K_N(t, s) \varphi(s) \, ds.$$

Then for any function $\varphi \in L^2[a, b]$ the value of $T_N \varphi$ will be represented in the form

$$T_N \varphi = \sum_{n, m=1}^{N} a_{nm} \psi_n(t) \int_a^b \psi_m(s) \varphi(s) \, ds$$

$$= \sum_{n=1}^{N} \left(\sum_{m=1}^{N} a_{nm} \int_a^b \psi_m(s) \varphi(s) \, ds \right) \cdot \psi_n(t) = \sum_{n=1}^{N} b_n \psi_n(t).$$

This means that the operator T_N maps the whole space $L^2[a, b]$ into a finite-dimensional space. Since in a finite-dimensional space any bounded set is compact, we conclude that T_N maps a bounded set into a compact set. Consequently, T_N is a completely continuous operator. Further, for any $\varphi \in L^2[a, b]$ we obtain the following estimate:

$$\| T_N \varphi - T \varphi \|^2 = \int_a^b \left[\int_a^b |K_N(t, s) - K(t, s)| \varphi(s) \, ds \right]^2 dt$$

$$\leqslant \int_a^b \int_a^b |K_N(t, s) - K(t, s)|^2 \, ds \int_a^b |\varphi(s)|^2 \, ds \, dt$$

$$= \| \varphi \|^2 \int_a^b \int_a^b |K_N(t, s) - K(t, s)|^2 \, dt \, ds.$$

Taking into account condition (8), we conclude that the relationship

$$\| T_N\varphi - T\varphi \| \to 0 \quad \forall \varphi \in L^2[a, b]$$

holds true, that is, the sequence of completely continuous operators $\{T_N\varphi\}$ converges uniformly to the operator $T\varphi$. Applying Theorem 2, we make sure that $T\varphi$ is completely continuous operator.

4. Representation of Completely Continuous Operators by Finite-dimensional Operators. Let us establish the connection between completely continuous and finite-dimensional operators.

Theorem 4. *Any completely continuous on the bounded set $M \subset E$ operator A is a uniform limit on M of the sequence $\{A_k\}$ of continuous finite-dimensional operators (which map M into a finite-dimensional subspace of the space E).*

The operator A is completely continuous, therefore the set $N = A(M)$ is compact. We choose a sequence of positive numbers $\{\varepsilon_k\}$, $\varepsilon_k \to 0$, and for each ε_k in the set N construct a finite ε-net:

$$N_k = \{y_1^{(k)}, y_2^{(k)}, \ldots, y_{m_k}^{(k)}\}, \quad y_\nu^{(k)} \in N, \quad k = 1, 2, \ldots .$$

Let us define on the set N the sequence of operators P_k:

$$P_k(y) = \frac{\sum_{l=1}^{m_k} \mu_l^{(k)}(y)\, y_l^{(k)}}{\sum_{l=1}^{m_k} \mu_l^{(k)}(y)}, \quad y \in N,$$

where

$$\mu_l^{(k)}(y) = \begin{cases} \varepsilon_k - \| y - y_l^{(k)} \| & \text{for } \| y - y_l^{(k)} \| \leqslant \varepsilon_k, \\ 0 & \text{for } \| y - y_l^{(k)} \| \geqslant \varepsilon_k. \end{cases}$$

It is clear that at least for one l we have a strict inequality $\mu_l^{(k)}(y) > 0$ (for any $y \in N$ there is $y_\nu^{(k)}$ such that $\| y - y_\nu^{(k)} \| < \varepsilon_k$). Since all $\mu_l^{(k)}(y)$ are continuous on N, the operator $P_k(y)$ is also continuous on N $\left(\sum_{l=1}^{m_k} \mu_l^{(k)}(y) > 0 \right.$ **for all** $\left. y \in N \right)$.

Further, the following inequality is valid:

$$\| y - P_k(y) \| = \left\| y - \frac{\sum_{i=1}^{m_k} \mu_i^{(k)}(y)\, y_i^{(k)}}{\sum_{i=1}^{m_k} \mu_i^{(k)}(y)} \right\| = \left\| \frac{\sum_{i=1}^{m_k} \mu_i^{(k)}(y)\,(y - y_i^{(k)})}{\sum_{i=1}^{m_k} \mu_i^{(k)}(y)} \right\| \leqslant \frac{\sum_{i=1}^{m_k} \mu_i^{(k)}(y) \| y - y_i^{(k)} \|}{\sum_{i=1}^{m_k} \mu_i^{(k)}(y)} < \varepsilon_k.$$

Here we have taken into consideration that if $\| y - y_i^{(k)} \| \geqslant \varepsilon_k$, then the approximate functions $\mu_i^{(k)}(y)$ are equal to zero, that is, $\mu_i^{(k)}(y) = 0$. Setting now $A_k(x) = P_k(Ax)$, for any $x \in M$ we derive the estimate

$$\| Ax - A_k x \| = \| Ax - P_k(Ax) \| < \varepsilon_k,$$

from which we conclude that the sequence of finite-dimensional operators A_k converges to the operator A.

Remark. The set of values of the operator $P_k(y)$ is contained in the convex hull generated by the elements of the ε-net. This fact will be taken advantage of later.

5. Compactness of the Range of Values of a Sequence of Completely Continuous Operators. Let us prove one more property of sequences of completely continuous operators.

Theorem 5. *Let the sequence of completely continuous on the set M operators $\{A_n\}_{n=1}^{\infty}$ converge uniformly on M to the operator A_0 and let $N_n = A_n(M)$, $n = 0, 1, \ldots$, be ranges of values of the operators A_n on the set M. Then the set $N = \bigcup_{n=0}^{\infty} N_n$ is compact.*

By Theorem 3, the operator A_0 is completely continuous. The uniform convergence of the sequence $\{A_n\}$ to the operator A_0 implies that $\forall \varepsilon > 0\ \exists\, n_0 = n_0(\varepsilon)$ such that $\forall y \in N_n$, $n > n_0$, there exists an element $u_y \in N_0$ for which the inequality $\| y - u_y \| < \varepsilon$ takes place. This follows from the fact that if x' is one of the preimages of the element y, i.e. $y = A_{n_0} x'$, then, as u_y, it is possible to

take $u' = A_0 x' \in N_0$ or the points lying in a certain neighbourhood of u'.

Consider the set $K = \bigcup_{n=0}^{n_0} N_n$ which is compact as the union of a finite number of compact sets. In the set N there is a finite ε-net for K. Let us show that it is also a finite ε-net for N. Indeed, if $y \in K$, then there is nothing to prove. And if $y \in N_n$ for $n > n_0$, then in the set N_0 we choose u_y such that $\| y - u_y \| < \varepsilon$. Here $u_y \in N_0 \subset K$, and therefore there is an element y_ε from the finite ε-net in K for which $\| y_\varepsilon - u_y \| < \varepsilon$. Then for this element y_ε we have the inequality $\| y - y_\varepsilon \| < 2\varepsilon$ and, since a finite ε-net is present in K, we make sure that N is compact.

Sec. 4.3. SCHAUDER'S THEOREM AND ITS APPLICATION

1. Schauder's Theorem. In Sec. 3.4 we proved the contraction mapping principle which establishes the existence and uniqueness of the fixed point of the contraction operator A. If the uniqueness of the fixed point is not required, then the condition of contractability of the operator A can be made somewhat weaker.

Theorem 1 (Schauder's Theorem). *If a completely continuous operator A maps the closed convex set $\bar{S}$ of the Banach space X into itself, then there is a fixed point of this mapping, that is, a point $x_0 \in \bar{S}$ such that $Ax_0 = x_0$.*

According to Theorem 4 from the preceding section, let us represent the operator A as a uniform limit on $\bar{S}$ of the sequence of finite-dimensional operators $\{A_n\}$. Since $\bar{S}$ is convex, for any $x \in \bar{S}$ the inclusion $A_n x \in \bar{S}$ is valid. This follows from the fact that the operators constructed in the above-mentioned theorem are contained in the linear hull of elements of the corresponding ε-net which is convex.

Let E_n be a finite-dimensional subspace containing $A_n(\bar{S})$. Consider the set $S_n = (\bar{S} \cap E_n)$ which is embedded in E_n. The set S_n is convex in n-dimensional subspace. It is clear that $A_n(\bar{S}) \subset \bar{S}$, $A_n(\bar{S}) \subset E_n$ and, all the more, $A_n(S_n) \subset S_n$. Thus, the operator A_n maps the convex set of the n-dimensional subspace into itself.

Let us now apply Brouwer's theorem proved in topology: *if a continuous finite-dimensional operator f maps a closed convex set $\bar{S}$ of the finite-dimensional space into itself, then there is in $\bar{S}$ a fixed point of the operator f.*

According to this theorem, there exists a point $x_n \in S_n$ such that $A_n x_n = x_n$. But $S_n \subset \bar{S}$, and therefore x_n is a fixed point of the operator A_n also when $\bar{S}$ is mapped into itself. For each point x_n the inclusion $x_n \in A_n(\bar{S})$ is valid, and therefore the following enclosure is also valid:

$$\{x_n\} \subset S^* = \bigcup_{n=1}^{\infty} A_n(\bar{S}) \subset \bar{S}.$$

By Theorem 5 proved in the preceding section the set S^* is compact. Therefore we separate from the sequence $\{x_n\}$ the convergent sequence $\{x_{n_\nu}\}$, whose limit x_0, by virtue of the closeness of $\bar{S}$, also belongs to $\bar{S}$.

Let us estimate the norm of the difference $Ax_0 - x_0$:

$$\begin{aligned} \| Ax_0 - x_0 \| &\leqslant \| Ax_0 - Ax_{n_\nu} \| + \| Ax_{n_\nu} - A_{n_\nu} x_{n_\nu} \| \\ &\quad + \| A_{n_\nu} x_{n_\nu} - x_0 \| = \| Ax_0 - Ax_{n_\nu} \| \\ &\quad + \| Ax_{n_\nu} - A_{n_\nu} x_{n_\nu} \| + \| x_{n_\nu} - x_0 \|. \end{aligned}$$

Given $\varepsilon > 0$, we choose $N_1(\varepsilon)$ so large that for $n_\nu > N_1(\varepsilon)$ the following inequalities are fulfilled:

$$\| x_{n_\nu} - x_0 \| < \frac{\varepsilon}{3} \quad \text{and} \quad \| Ax_0 - Ax_{n_\nu} \| < \frac{\varepsilon}{3},$$

the first of these inequalities follows from the fact that $\lim_{n_\nu \to \infty} x_{n_\nu} = x_0$, and the second follows from the continuity of the operator A. Let us now choose $N_2(\varepsilon)$ so large that for $n_\nu > N_2(\varepsilon)$ for all $x \in \bar{S}$ (and for $x = x_{n_\nu}$, in particular) we have the following inequality:

$$\| Ax - A_{n_\nu} x \| < \frac{\varepsilon}{3}.$$

This inequality is a consequence of the uniform convergence of the sequence $\{A_n\}$. Choosing $N = \max\{N_1(\varepsilon), N_2(\varepsilon)\}$, we obtain that for $n_\nu > N$ the following estimate is true:

$$\| Ax_0 - x_0 \| < \varepsilon.$$

But the left-hand side of this inequality is a constant number, consequently, $Ax_0 = x_0$, and this means that x_0 is a fixed point of the operator A.

2. Existence of Solution of a Differential Equation. Let us apply Schauder's theorem to the proof of the existence of a solution of a second-order differential equation.

Example 1°. Show that if the function $f(t, x, dx/dt)$ is continuous with respect to its arguments and is bounded in the domain

$$D = \begin{cases} 0 \leqslant t \leqslant T, \\ -\infty < x < \infty, \\ -\infty < \dfrac{dx}{dt} < \infty, \end{cases}$$

then for $0 \leqslant t \leqslant T$ there is the solution of the equation

$$\frac{d^2x}{dt^2} = f\left(t,\ x(t),\ \frac{dx}{dt}\right) \tag{1}$$

satisfying the boundary conditions

$$x(0) = x_1, \quad x(T) = x_2. \tag{2}$$

Consider the function

$$G(t, s) = \begin{cases} \dfrac{s(T-t)}{T} \text{ for } 0 \leqslant s \leqslant t \leqslant T, \\ \dfrac{t(T-s)}{T} \text{ for } 0 \leqslant t \leqslant s \leqslant T, \end{cases} \tag{3}$$

and show that any solution of the integral equation

$$x(t) = x_1 + \frac{x_2 - x_1}{T}\, t - \int_0^T G(t, s) f\left(s,\ x(s),\ \frac{dx(s)}{ds}\right) ds \tag{4}$$

is the solution of equation (1) satisfying conditions (2).

Indeed, the fulfilment of conditions (2) is obvious since $G(0, s) = G(T, s) = 0$. Breaking integral (4) into the intervals $0 \leqslant s \leqslant t$ and $t \leqslant s \leqslant T$ and differentiating the obtained integrals with respect to the parameter t and to

the variable upper and lower limits, we obtain the equalities

$$x'(t) = \frac{x_2 - x_1}{T} - \int_0^T \frac{s}{T} f(s, x, x')\,ds - \frac{t(T-t)}{T} f(t, x(t), x'(t)) - \int_t^T \frac{T-s}{T} f(s, x, x')\,ds + \frac{t(T-t)}{T} f(t, x(t), x'(t))$$

$$= \frac{x_2 - x_1}{T} + \int_0^T \frac{s}{T} f(s, x, x')\,ds - \int_t^T f(s, x, x')\,ds \quad (5)$$

and

$$x''(t) = f(t, x(t), x'(t)).$$

The last equality means that if the function $x(t)$ is the solution of equation (4), then it also satisfies equation (1). Therefore, it suffices to show the solvability of equation (4).

Let E be a Banach space of continuously differentiable functions $x(t)$, $0 \leqslant t \leqslant T$, with the norm

$$\|x\| = \sup_{0 \leqslant t \leqslant T} \{|x(t)| + |x'(t)|\}.$$

Consider the operator A whose value for each function $x(t) \in E$ is written in the form

$$Ax \equiv x_1 + \frac{x_2 - x_1}{T} t - \int_0^T G(t, s) f(s, x(s), x'(s))\,ds.$$

By analogy with the derivation of equality (5), we see that the function $Ax(t)$ is also continuously differentiable, that is, the operator A maps the space E into itself. The continuity of the operator A is an obvious consequence of the continuity of $f(t, x, x')$.

Let us prove that the image $A(E)$ is a compact. Indeed, let $|f| \leqslant M$ in the domain D. Then taking into account the estimate

$$0 \leqslant G(t, s) \leqslant \frac{T}{4}, \quad 0 \leqslant s, \quad t \leqslant T,$$

following from relationship (3), we get the inequality

$$|Ax(t)| \leqslant |x_1| + |x_2 - x_1| + \frac{MT^2}{4}. \quad (6)$$

With the aid of the transformation performed when deriving relationship (5), we get the equality

$$(Ax)'_t = \frac{x_2 - x_1}{T} + \int_0^T \frac{s}{T} f(s, x, x')\,ds - \int_t^T f(s, x, x')\,ds, \quad (7)$$

from which there follows the estimate

$$|(Ax)'| \leqslant \frac{|x_2 - x_1|}{T} + 2MT. \quad (8)$$

From inequalities (6) and (8) we conclude that the family of functions $\{Ax\}$, $x \in E$, is equibounded.

Assuming for definiteness that $t_1 > t_2$ from equation (4) we find the relationship

$$Ax(t_1) - Ax(t_2) = \frac{x_2 - x_1}{T}(t_1 - t_2) - \int_0^T [G(t_1, s) - G(t_2, s)] f(s, x, x')\,ds.$$

Taking into consideration the equality

$$G(t_1, s) - G(t_2, s) = \begin{cases} 0 \text{ for } 0 \leqslant s \leqslant t_2 \text{ and for } t_1 \leqslant s \leqslant T, \\ \dfrac{t_2(T-s)}{T} - \dfrac{s(T-t_1)}{T} \text{ for } t_2 \leqslant s \leqslant t_1, \end{cases}$$

we get the following representation:

$$Ax(t_1) - Ax(t_2) = \frac{x_2 - x_1}{T}(t_1 - t_2) + \int_{t_2}^{t_1} (s - t_2) f(s, x, x')\,ds - \int_{t_2}^{t_1} \frac{s}{T}(t_1 - t_2) f(s, x, x')\,ds.$$

Bearing in mind that $|f| \leqslant M$, we derive the estimate

$$|Ax(t_1) - Ax(t_2)| \leqslant \frac{|x_2 - x_1|}{T} |t_1 - t_2| + 2M(t_1 - t_2)^2. \quad (9)$$

Besides, from expression (7) there follows the equality

$$(Ax)'(t_1) - (Ax)'(t_2) = -\int_{t_2}^{t_1} f(s, x, x')\,ds,$$

whence we get the inequality

$$|(Ax)'(t_1) - (Ax)'(t_2)| \leqslant M|t_1 - t_2|, \qquad (10)$$

and from estimates (9) and (10) we make sure that the family of functions $\{Ax\}$, $x \in E$, is equicontinuous with respect to the norm of the space E. Applying Arzela's theorem to the space E (we shall not dwell on its proof), we make sure that the set $A(E)$ is compact in E.

Thus, the operator A is completely continuous and, by Schauder's theorem, there exists a fixed point of this operator, and this means that equation (4) has a solution, and therefore equation (1) with condition (2) also has a solution.

3. Schauder's Theorem Applied to Monotone Operators. Schauder's theorem is frequently used when applying iterative methods to monotone operators. To illustrate this application, let us first introduce the following definitions.

A set M is said to be *partially ordered* if for some pairs (x and y) of elements of this set an *order relation* is introduced (for instance, x precedes y). This order relation will be written in the form $x \prec y$; $x \in M$, $y \in M$.

And if the order relation $\prec$ is introduced for any pairs x, y of elements of the set M, then the set M will be termed *completely ordered.*

For example, the set of infinite dimensional vectors $x = (x_1, x_2, \ldots, x_n, \ldots)$ becomes completely ordered if the order relation $\prec$ is introduced in the following way: $x \prec y$ if $x_\nu < y_\nu$ for $\nu = 1, 2, \ldots, m$ and $x_{m+1} < y_{m+1}$.

In the space $C[a, b]$, it is possible to introduce a partial ordering, assuming that $x(t) \prec y(t)$ if $x(t) \leqslant y(t)$ for all $t \in [a, b]$.

The operator A: $(X \to X)$ mapping the space X into itself is said to be *monotonically increasing* if the relation $x \prec y$ ($x, y \in X$) implies an analogous relation of images: $Ax \prec Ay$. And if the relation $x \prec y$ implies the relation $Ay \prec Ax$, then the operator A is said to be *monotonically decreasing.*

The existence of such operators is readily confirmed by the following examples.

Example 2°. Let on the half-ordered space $C[0, 1]$ there be given the operator

$$Tx = \int_a^b K(s, t, x(t))\,dt,$$

whose kernel $K(s, t, x(t))$ for $s, t \in [a, b]$ and any $x \in C[0, 1]$ satisfies the condition $\partial K/\partial x \geqslant 0$. Prove that the operator T is monotonically increasing.

Indeed, if $x(t) \prec y(t)$, that is, we have $x(t) \leqslant y(t)$ for $0 \leqslant t \leqslant 1$; then, by setting

$$\tilde{x}(t) = x(t) + \theta(t)[y(t) - x(t)], \quad 0 \leqslant \theta(t) \leqslant 1,$$
$$\tilde{x}(t) \in C[0, 1],$$

we obtain the relationship

$$Ty - Tx = \int_a^b [K(s, t, y(t)) - K(s, t, x(t))]\,dt$$
$$= \int_a^b \frac{\partial K(s, t, \tilde{x}(t))}{\partial x}[y(t) - x(t)]\,dt \geqslant 0,$$

and this means that $Tx \prec Ty$.

Example 3°. Let on a completely ordered n-dimensional Euclidean space E_n there be given a stretching operator A defined by the matrix

$$A = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$

Prove that if $\lambda_\nu > 0$ for $\nu = 1, 2, \dots, n$, then the operator A is monotonically increasing.

Indeed, if

$$x = (x_1, \dots, x_n) \prec y = (y_1, \dots, y_n),$$

then

$$x_j = y_j, \quad j = 1, 2, \dots, m, \quad \text{and} \quad x_{m+1} < y_{m+1}, \quad m < n.$$

But then

$\lambda_j x_j = \lambda_j y_j$, $j = 1, 2, \ldots, m$ and $\lambda_{m+1} x_{m+1} < \lambda_{m+1} y_{m+1}$,

i.e. $Ax \prec Ay$.

The operator A: $(X \to X)$ is called *linear* if for any numbers α and β and any elements x and y from X there is the relationship $A(\alpha x + \beta y) = \alpha Ax + \beta Ay$.

Theorem 2. *Let in a partially ordered Banach space X there be given the equation $Tx \equiv Ax + f = x$, where A is a completely continuous linear operator representable in the form of the sum of a monotonically increasing operator A_1 and a monotonically decreasing operator A_2 defined and continuous in the convex domain $K \subset X$. Suppose an iteration is carried out using the formulas*

$$y_{n+1} = A_1 y_n + A_2 z_n + f, \quad z_{n+1} = A_1 z_n + A_2 y_n + f,$$
$$n = 0, 1, \ldots,$$

beginning with the elements $y_0 \in K$ and $z_0 \in K$, where for the elements y_0, z_0, y_1, z_1 the order relation $y_0 \prec y_1 \prec z_1 \prec z_0$ is valid. Then the operator T maps the interval $M_n = [y_n, z_n]$, that is, the set of points $x = \alpha y_n + (1 - \alpha) z_n$, $0 \leqslant \alpha \leqslant 1$, into itself, and the equation $Tx = x$ has a solution x such that $y_n \prec \tilde{x} \prec z_n$, $n = 0, 1$.

Let us prove by induction that for all $n = 0, 1, \ldots$ the following order relation is fulfilled:

$$y_n \prec y_{n+1} \prec z_{n+1} \prec z_n. \tag{11}$$

Let $y_{n-1} \prec y_n \prec z_n \prec z_{n-1}$. Then the following relationships are valid:

$$A_1 y_n \prec A_1 z_n \quad \text{and} \quad A_2 z_n \prec A_2 y_n.$$

Applying these relationships, we derive the order relation

$$y_{n+1} = A_1 y_n + A_2 z_n + f \prec A_1 z_n + A_2 y_n + f = z_{n+1}.$$

From the relationships $A_1 y_{n-1} \leqslant A_1 y_n$ and $A_2 z_{n-1} \leqslant A_2 z_n$ there follows the validity of the order relations

$$y_n = A_1 y_{n-1} + A_2 z_{n-1} + f \prec A_1 y_n + A_2 z_n + f = y_{n+1}$$

and

$$z_{n+1} = A_1 z_n + A_2 y_n + f \prec A_1 z_{n-1} + A_2 y_{n-1} + f = z_n.$$

Thus, relationships (11) have been proved.

From the convexity of the domain K it follows that the interval $M_n = [y_n, z_n]$ wholly belongs to K, and, by virtue of (11), is mapped by the completely continuous operator T into itself: $T(M_n) = M_{n+1} \subseteq M_n$. Taking into consideration the continuity of the operators A_1 and A_2 and the relationships

$$y_{n+1} = A_1 y_n + A_2 z_n + f, \quad z_{n+1} = A_1 z_n + A_2 y_n + f,$$

as $n \to \infty$, we obtain the equalities

$$y = A_1 y + A_2 z + f, \quad z = A_1 z + A_2 y + f,$$

in which $y \prec z$. Setting $\tilde{x} = (1/2)(y + z)$, we have the relationship $y_n \prec \tilde{x} \prec z_n$ and the equality $\tilde{x} = A\tilde{x} + f$, and this means that $\tilde{x}$ is the solution of the original equation.

Example 4°. In a partially ordered space of doubly differentiable functions $C^2[0, 1]$, establish the boundaries for solving the equation

$$\frac{d^2x}{dt^2} = -t - \sqrt{x} \qquad (12)$$

with the boundary conditions

$$x(0) = 0, \quad x(1) = 1. \qquad (13)$$

Equation (12) with conditions (13) is equivalent to the integral equation

$$x(t) = t + \int_0^1 G(t, s)(s + \sqrt{x(s)})\, ds,$$

where

$$G(t, s) = \begin{cases} s(1-t) & \text{for } 0 \leqslant s \leqslant t \leqslant 1, \\ t(1-s) & \text{for } 0 \leqslant t \leqslant s \leqslant 1. \end{cases}$$

If we choose $y_0(t) = t^2$, $z_0(t) = (2\sqrt{t} - t)^2$, then from the integral equation we shall find the first approximations:

$$y_1(t) = t + \int_0^1 G(t, s)\, 2s\, ds = \frac{4}{3}t - \frac{t^3}{3}$$

and

$$z_1(t) = t + \int_0^1 G(t, s)\, 2\sqrt{s}\, ds = \frac{23}{15}t - \frac{8}{15}t^{5/2}.$$

It is not difficult to check that the following order relations are fulfilled:

$$y_0(t) \prec y_1(t) \prec z_1(t) \prec z_0(t),$$

therefore for the solution $x(t)$ of equation (12) we have the estimate

$$\frac{4}{3}t - \frac{t^3}{3} \prec x(t) \prec \frac{23}{15}t - \frac{8}{15}t^{5/2}.$$

Sec. 4.4. ITERATION METHOD FOR SOLVING FREDHOLM'S EQUATION

1. Properties of Iterated Fredholm's Operators. Let us first establish certain properties of *Fredholm's iterated kernels* and *operators*. The same as in Sec. 4.2, we shall assume that the kernel $K(t, s)$ is quadratically summable in the square $a \leqslant t \leqslant b$, $a \leqslant s \leqslant b$, that is, satisfies the condition

$$B^2 = \int_a^b \int_a^b |K(t, s)|^2\, dt\, ds < \infty. \tag{1}$$

It was shown in Subsection 3 of Sec. 4.2 that for every function $\varphi(t) \in L^2[a, b]$ the value of Fredholm's operator

$$T\varphi = \int_a^b K(t, s)\, \varphi(s)\, ds \tag{2}$$

again belongs to $L^2[a, b]$. Hence it follows that the operator T can also be applied to the value $T\varphi$. Then we have

the representation

$$TT\varphi = T^2\varphi = \int_a^b K(t,\ s) \int_a^b K(s,\ s_1)\,\varphi(s_1)\,ds_1\,ds$$

$$= \int_a^b \left(\int_a^b K(t,\ s)\,K(s,\ s_1)\,ds \right) \varphi(s_1)\,ds_1,$$

whence we obtain the form of *Fredholm's second iterated kernel*

$$K_2(t,\ s) = \int_a^b K(t,\ s_1)\,K(s_1,\ s)\,ds_1.$$

Let us estimate the norm of the element $T^2\varphi$ which also belongs to $L^2[a,\ b]$. Applying the Cauchy-Buniakowski inequality, we get the estimate

$$\| T^2\varphi \|^2 = \int_a^b | T^2\varphi |^2\,dt$$

$$= \int_a^b \left| \int_a^b \left(\int_a^b K(t,\ s)\,K(s,\ s_1)\,ds \right) \varphi(s_1)\,ds_1 \right|^2 dt$$

$$\leqslant \int_a^b \left(\int_a^b | \varphi(s_1) |^2\,ds_1 \int_a^b \left| \int_a^b K(t,\ s)\,K(s,\ s_1)\,ds \right|^2 ds_1 \right) dt$$

$$\leqslant \| \varphi \|^2 \int_a^b \int_a^b \left(\int_a^b | K(t,\ s) |^2\,ds \int_a^b | K(s,\ s_1) |^2\,ds \right) ds_1\,dt$$

$$= \| \varphi \|^2 \left(\int_a^b \int_a^b | K(t,\ s) |^2\,dt\,ds \right)$$

$$\times \left(\int_a^b \int_a^b | K(s,\ s_1) |^2\,ds\,ds_1 \right) = B^4 \| \varphi \|^2.$$

Hence we derive the inequality

$$\| T^2\varphi \| \leqslant B^2 \| \varphi \|, \tag{3}$$

from which we conclude that for the norm of the operator T^2 the following estimate holds true:

$$\| T^2 \|_1 = \sup_{\| \varphi \| \leqslant 1} \| T^2 \varphi \| \leqslant B^2.$$

It is possible to show in a similar way that for the *nth Fredholm iterated kernel* we have the following representation:

$$K_n(t, s) = \int_a^b \int_a^b \ldots \int_a^b K(t, s_1) K(s_1, s_2) \ldots$$

$$\ldots K(s_{n-1}, s)\, ds_1\, ds_2 \ldots ds_{n-1},$$

and for the norm of the nth iterated operator T^n the following estimate holds true:

$$\| T^n \|_1 \leqslant B^n. \tag{4}$$

2. Solving Fredholm's Integral Equation. Let us now pass to solving the so-called *Fredholm's integral equation*

$$\varphi(t) - \lambda \int_a^b K(t, s) \varphi(s)\, ds = f(t), \tag{5}$$

that is, the equation $\varphi - \lambda T \varphi = f$.

Taking $\varphi_0 = f$ for the initial approximation, we have the following equalities:

$$\varphi_1 = f + \lambda T \varphi_0 = f + \lambda T f,$$
$$\varphi_2 = f + \lambda T \varphi_1 = f + \lambda T f + \lambda^2 T^2 f,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\varphi_n = f + \lambda T f + \ldots + \lambda^n T^n f.$$

The function φ_n can be regarded as the partial sum of the series

$$\sum_{m=0}^{\infty} \lambda^m T^m f, \tag{6}$$

called *Neumann's series.*

Theorem 1. *If the kernel $K(t, s)$ of equation (5) is quadratically summable on the square $a \leqslant t, s \leqslant b$, and $|\lambda| <$*

$< 1/B$, *where*

$$B^2 = \int_a^b \int_a^b |K(t, s)|^2 \, dt \, ds,$$

then Neumann's series (6) *converges in the mean to the square summable solution of equation* (5), *this solution being unique.*

Let us estimate the quantity

$$\| \varphi_{n+p} - \varphi_n \| = \Big\| \sum_{m=n+1}^{n+p} \lambda^m T^m f \Big\|.$$

Applying inequalities (3) and (4), we obtain the relationship

$$\| \varphi_{n+p} - \varphi_n \| \leqslant \sum_{m=n+1}^{n+p} |\lambda|^m \| T^m f \| \leqslant \sum_{m=n+1}^{n+p} |\lambda|^m \| T^m \|_1 \| f \|$$

$$\leqslant \| f \| \sum_{m=n+1}^{n+p} (|\lambda| B)^m \leqslant \| f \| \frac{(|\lambda| B)^{n+1}}{1 - |\lambda| B}.$$

If $|\lambda| < 1/B$, then the right-hand member tends to zero as $n \to \infty$, and this means that $\{\varphi_n\}$ is a Cauchy sequence. By virtue of the completeness of the space $L^2[a, b]$, there exists in it a function φ^* such that

$$\lim_{n \to \infty} \varphi_n = \varphi^*, \quad \text{i.e.} \quad \| \varphi_n - \varphi^* \| \to 0 \text{ as } n \to \infty.$$

Let us prove that φ^* is the solution of equation (5). Indeed, since $\varphi_n = f + \lambda T \varphi_{n-1}$, as $n \to \infty$, the left-hand member of this equality tends to φ^*, and since, by inequality (3), the following relationship holds true:

$$\| T\varphi_{n-1} - T\varphi^* \| = \| T(\varphi_{n-1} - \varphi^*) \|$$

$$\leqslant B \| \varphi_{n-1} - \varphi^* \| \to 0,$$

the right-hand side tends to the expression

$$f + \lambda T \varphi^*.$$

This means that φ^* is the solution of equation (5).

Suppose that there are two solutions of equation (5): $\varphi^* \in L^2[a, b]$ and $\varphi^{**} \in L^2[a, b]$. Then the equalities

below are valid:

$$\varphi^* = f + \lambda T\varphi^* \quad \text{and} \quad \varphi^{**} = f + \lambda T\varphi^{**},$$

from which we derive the relationship

$$(\varphi^* - \varphi^{**}) = \lambda T\,(\varphi^* - \varphi^{**}).$$

Applying estimate (3), we get the inequality

$$\| \varphi^* - \varphi^{**} \| = | \lambda | \, \| T\,(\varphi^* - \varphi^{**}) \| \leqslant | \lambda | \, B \, \| \varphi^* - \varphi^{**} \|.$$

Taking into account that $| \lambda | < 1/B$, we obtain an impossible relationship

$$\| \varphi^* - \varphi^{**} \| < \| \varphi^* - \varphi^{**} \|.$$

Thus, our supposition on the existence of two solutions leads to a contradiction; hence, $\varphi^* = \varphi^{**}$.

Remark. If $| \lambda | \geqslant 1/B$, then, depending on the kernel $K(t, s)$, the solution of equation (5) may, nevertheless, exist. The following example makes this sure.

Example. Show that the equation

$$\varphi(s) - \lambda \int_0^1 \varphi(s)\,ds = 1 \tag{7}$$

has a solution for both $| \lambda | < 1$ and $| \lambda | > 1$.

The kernel of equation (7) is identically equal to unity, and then, by Theorem 1, the solution of this equation for $| \lambda | < 1$ does exist. Let us find this solution. Choosing $\varphi_0(x) \equiv 1$, we have the equalities

$$\varphi_1(x) = 1 + \lambda \int_0^1 1\,ds = 1 + \lambda,$$

$$\varphi_2(x) = 1 + \lambda \int_0^1 (1 + \lambda)\,ds = 1 + \lambda + \lambda^2,$$

. . . . . .

$$\varphi_n(x) = 1 + \lambda \int_0^1 (1 + \lambda + \ldots + \lambda^{n-1})\,ds = 1 + \lambda + \ldots + \lambda^n,$$

whence it follows that

$$\varphi_n(x) \to \frac{1}{1-\lambda} \quad \text{for} \quad |\lambda| < 1.$$

Thus, the solution of equation (7) is the function

$$\varphi(x) = \frac{1}{1-\lambda} \quad \text{for} \quad |\lambda| < 1.$$

On the other hand, noting that the integral $\int_0^1 \varphi(s)\, ds = C$ contained in equation (7) is constant, we arrive at the equality

$$\varphi(x) = 1 + \lambda C.$$

Integrating this equality between the limits from 0 to 1, we get the relationship

$$\int_0^1 \varphi(x)\, dx = 1 + \lambda C,$$

which can be written in the form

$$C = 1 + \lambda C.$$

Hence, for $\lambda \neq 1$ we find the value of the integral, that is, the constant $C = 1/(1 - \lambda)$. Consequently, for all $\lambda \neq 1$ the solution of equation (7) is the function

$$\varphi(x) = 1 + \frac{\lambda}{1-\lambda} = \frac{1}{1-\lambda}.$$

And if $\lambda = 1$, then the original equation (7) has no solution.

CHAPTER 5

Self-adjoint Operators in a Hilbert Space

Sec. 5.1. BASIC CONCEPTS OF A HILBERT SPACE

1. Scalar Product and Its Properties. A *Hilbert space* H is defined as a complete linear space over the field of complex numbers K in which every pair of elements x and y, usually called *vectors*, has an associated (generally speaking) complex number, which is called the *scalar (inner) product of the vectors*, denoted by (x, y) and satisfying the following conditions:

(a) $(y, x) = \overline{(x, y)}$;

(b) $(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 (x_1, y) + \alpha_2 (x_2, y)$, $\alpha_1, \alpha_2 \in K$;

(c) $(x, x) \geqslant 0$;

(d) the relationship $(x, x) = 0$ implies that $x = \theta$ (zero element of H).

The Euclidean space E_n, the functional space $L^2[a, b]$, and the coordinate space l^2, all of which we considered above, are particular cases of a Hilbert space. The scalar products in these spaces are defined in the following way:

if $x = (x_1, \ldots, x_n) \in E_n$, $y = (y_1, \ldots, y_n) \in E_n$, then $(x, y) = \sum_{k=1}^{n} x_k \overline{y_k}$;

if $f(x) \in L^2[a, b]$, $g(x) \in L^2[a, b]$, then $(f, g) = \int_a^b f(x)\overline{g(x)}\, dx$;

if $x = (x_1, \ldots, x_n, \ldots) \in l^2$, $y = (y_1, \ldots, y_n, \ldots) \in l^2$, then $(x, y) = \sum_{k=1}^{\infty} x_k \overline{y_k}$.

Scalar products possess the following properties:

(1) *If* $x \in H$, *then* $(\alpha x, \alpha x) = |\alpha^2| (x, x)$ *for any* $\alpha \in K$.

Using conditions (b), (a), and (c), we find

$$(\alpha x, \alpha x) = \alpha (x, \alpha x) = \alpha \overline{(\alpha x, x)} = \alpha\bar{\alpha} \overline{(x, x)} = |\alpha|^2 (x, x).$$

(2) *For any* x *and* y *from* H *Schwarz's inequality is true*:

$|(x, y)| \leqslant \sqrt{(x, x)}\sqrt{(y, y)}$.

For any $\lambda \neq 0$ and $y \neq 0$ we have the inequality $(x + \lambda y, x + \lambda y) \geqslant 0$. Hence we derive the relationship

$$(x, x) + \bar{\lambda} (x, y) + \lambda (y, x) + |\lambda|^2 (y, y) \geqslant 0. \qquad (1)$$

If we put $\lambda = -(x, y)/(y, y)$, then from relationship (1) it is possible to obtain the inequality

$$(x, x) - \frac{|(x, y)|^2}{(y, y)} \geqslant 0,$$

from which Schwarz's inequality follows.

(3) *For any* x *and* y *from* H *the triangle inequality*

$$\sqrt{(x+y, x+y)} \leqslant \sqrt{(x, x)} + \sqrt{(y, y)} \qquad (2)$$

holds true.

Revealing the scalar product and taking into consideration the inequality $z + \bar{z} \leqslant 2|z|$, which is true for complex numbers, we have the estimate

$$(x + y, x + y) = (x, x) + (x, y) + (y, x) + (y, y)$$
$$\leqslant (x, x) + (y, y) + 2|(x, y)|.$$

Using Schwarz's inequality, we derive the inequality

$$(x + y, x + y) \leqslant (x, x) + (y, y) + 2\sqrt{(x, x)}\sqrt{(y, y)}$$
$$= (\sqrt{(x, x)} + \sqrt{(y, y)})^2.$$

(4) *The following equality is obvious*:

$$(x + y, x + y) + (x - y, x - y) = 2 (x, x) + 2 (y, y);$$

it is called the parallelogram equality.

2. Norm in H. With the aid of a scalar product it is also possible to introduce the concept of norm in H, that is, to make H a normed linear space. We set $\|x\|^2 = (x, x)$.

Then the fulfilment of all properties of the norm is obvious. Property (4) of scalar products can be rewritten in terms of the norm, i.e.

$$\| x + y \|^2 + \| x - y \|^2 = 2 \| x \|^2 + 2 \| y \|^2,$$

and this is the geometrical theorem on the sum of the squared diagonals of a parallelogram generalized for the case of a Hilbert space.

The following theorem is also valid.

Theorem 1. *A scalar product is a continuous function with respect to convergence by norm, that is, if* $\| x_n - x \| \to 0$ *and* $\| y_n - y \| \to 0$ *as* $n \to \infty$, *then also*

$$| (x_n, y_n) - (x, y) | \to 0 \quad as \quad n \to \infty. \tag{3}$$

From the properties of scalar products we have the estimate

$$\begin{aligned} | (x_n, y_n) - (x, y) | &= | (x_n, y_n) - (x_n, y) \\ &\quad + (x_n, y) - (x, y) | \leqslant | (x_n, (y_n - y)) | \\ &\quad + | ((x_n - x), y) | \leqslant \sqrt{(x_n, x_n)}\sqrt{(y_n - y, y_n - y)} \\ &\quad + \sqrt{(x_n - x, x_n - x)}\sqrt{(y, y)} \\ &= \| x_n \| \| y_n - y \| + \| x_n - x \| \| y \|. \end{aligned}$$

The norms $\| x_n \|$ and $\| y \|$ are bounded. Therefore relationship (3) follows from this estimate.

Sec. 5.2. SELF-ADJOINT OPERATORS AND THEIR PROPERTIES

1. Definitions. In applications, an important role is played by self-adjoint operators, which we have to define now. Let an operator T mapping D into H be defined on a dense set D in a space H. The operator T is said to be *bounded* on D if there is a constant $M > 0$ such that for all $x \in D$ the following estimate is valid:

$$\| Tx \| \leqslant M \| x \|. \tag{1}$$

The least of such constants M is called the *norm of the operator* T on D and is denoted by $\| T \|_1$. Then inequality (1)

can be written in the form

$$\| Tx \| \leqslant \| T \|_1 \| x \|. \tag{2}$$

The norm of the operator T can also be defined with the aid of the relationship

$$\| T \|_1 = \sup_{x \in D} \frac{\| Tx \|}{\| x \|} = \sup_{\|x\| \leqslant 1} \| Tx \|,$$

but we are not going to dwell on its proof.

The operator T is said to be *self-adjoint* on H if for any elements x and y from H the relationship $(Tx, y) = (x, Ty)$ holds true.

Let us give a few examples of self-adjoint operators.

Example 1°. Show that the averaging operator $Tf = (1/2)[f(x+s) + f(x-s)]$ is self-adjoint when considered in the space $L^2(-\infty, \infty)$.

Let us verify the condition of self-adjointness of this operator:

$$(Tf, g) = \int_{-\infty}^{\infty} \frac{1}{2} [f(x+s) + f(x-s)] \overline{g}(x)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{2} [\overline{g}(u-s) + \overline{g}(u+s)] f(u)\, du = (f, Tg).$$

Thus the averaging operator is self-adjoint in the space $L^2(-\infty, \infty)$.

When solving the equations of mathematical physics by using the net method, partial derivatives are replaced by difference relationships, for instance, $\partial f/\partial x$ is replaced by either the relationship $\frac{1}{h}\left[f\left(x+\frac{h}{2}\right) - f\left(x-\frac{h}{2}\right)\right]$ or $\frac{1}{h}[f(x+h) - f(x)]$, while the second derivative $\partial^2 f/\partial x\, \partial y$ can be replaced by the relationship

$$\frac{1}{h^2}\left[f\left(x+\frac{h}{2}, y+\frac{h}{2}\right) - f\left(x-\frac{h}{2}, y+\frac{h}{2}\right)\right.$$
$$\left. - f\left(x+\frac{h}{2}, y-\frac{h}{2}\right) + f\left(x-\frac{h}{2}, y-\frac{h}{2}\right)\right]$$

or the relationship

$$\frac{1}{h^2}[f(x+h,\ y+h)-f(x,\ y+h)-f(x+h,\ y)+f(x,\ y)].$$

Therefore it is of interest to study in the space $L^2(-\infty,\ \infty)$ the operators

$$Tf=i\left[f\left(x+\frac{h}{2}\right)-f\left(x-\frac{h}{2}\right)\right],$$

$$Sf=i[f(x+h)-f(x)],$$

$$Uf=f\left(x+\frac{h}{2},\ y+\frac{h}{2}\right)-f\left(x-\frac{h}{2},\ y+\frac{h}{2}\right)$$
$$-f\left(x+\frac{h}{2},\ y-\frac{h}{2}\right)+f\left(x-\frac{h}{2},\ y-\frac{h}{2}\right),$$

and

$$Vf=f(x+h,\ y+h)-f(x,\ y+h)$$
$$-f(x+h,\ y)+f(x,\ y).$$

Example 2°. Show that the operator

$$Tf=i\left[f\left(x+\frac{h}{2}\right)-f\left(x-\frac{h}{2}\right)\right]$$

is self-adjoint in the space $L^2(-\infty,\ \infty)$.

We have the relationships

$$(Tf,\ g)=\int_{-\infty}^{\infty} i\left[f\left(x+\frac{h}{2}\right)-f\left(x-\frac{h}{2}\right)\right]\overline{g}(x)\,dx$$

$$\int_{\infty}^{\infty} if(u)\left[\overline{g}\left(u-\frac{h}{2}\right)-\overline{g}\left(u+\frac{h}{2}\right)\right]du$$

$$\int_{\infty}^{\infty} f(u)\left\{\overline{i\left[g\left(u+\frac{h}{2}\right)-g\left(u-\frac{h}{2}\right)\right]}\right\}du,$$

that is, in fact, $(Tf,\ g)=(f,\ Tg)$.

It can also be readily verified that the operator U is self-adjoint and the operators S and V are not.

Example 3°. Find the condition of self-adjointness of Fredholm's operator

$$T\varphi = \int_a^b K(t, s)\,\varphi(s)\,ds. \tag{3}$$

For any $\varphi \in L^2[a, b]$ and $\psi \in L^2[a, b]$ we have the equalities

$$(T\varphi, \psi) = \int_a^b \left(\int_a^b K(t, s)\,\varphi(s)\,ds \right) \overline{\psi}(t)\,dt$$

$$= \int_a^b \varphi(s) \left(\int_a^b K(t, s)\,\overline{\psi}(t)\,dt \right) ds$$

$$= \int_a^b \varphi(s) \left(\int_a^b \overline{K(t, s)}\,\psi(t)\,dt \right) ds = (\varphi, \mu),$$

where $\mu(t) = \int_a^b \overline{K(s, t)}\,\psi(s)\,ds$.

Hence we conclude that if $\overline{K(s, t)} = K(t, s)$, the *Fredholm's operator* defined by relationship (3) will be *self-adjoint*; in the case of a real kernel $K(t, s)$, the condition of its self-adjointness consists in symmetry, that is, a real Fredholm operator is self-adjoint if $K(s, t) = K(t, s)$.

2. Computing the Norm of a Self-adjoint Operator. Let us prove the lemma on the connection between the norm of an operator and a scalar product.

Lemma 1. *For the norm of any self-adjoint operator T mapping a space H into H we have the equation*

$$\| T \|_1 = \sup_{\|x\| \leqslant 1} \| Tx \| = \sup_{\|x\|=1} |(Tx, x)|.$$

Applying Schwarz's inequality and taking into consideration inequality (2), we have the estimate

$$|(T\varphi, \varphi)| \leqslant \| T\varphi \| \| \varphi \| \leqslant \| T \|_1 \| \varphi \|^2.$$

Hence there follows the inequality

$$|(T\varphi, \varphi)| \leqslant C_T \| \varphi \|^2, \tag{4}$$

the following estimate being valid for the constant C_T:

$$C_T \leqslant \| T \|_1. \tag{5}$$

Let now $z \neq 0$ be an arbitrary element from H. Setting $u = (1/\alpha)\, Tz$, where $\alpha = (\| Tz \|/\| z \|)^{1/2}$, we estimate the quantity from above

$$\| Tz \|^2 = (Tz,\ Tz) = (Tz,\ \alpha u).$$

Using properties (a) and (b) of scalar products and the self-adjointness of the operator T, we may write the equalities

$$(Tz,\ \alpha u) = \overline{(\alpha u,\ Tz)} = \alpha\, \overline{(u,\ Tz)} = \alpha\, (Tz,\ u)$$
$$= \alpha\, (z,\ Tu) = (\alpha z,\ Tu),$$

and this means that $(Tz,\ \alpha u) = (T\alpha z,\ u)$. Applying the properties of scalar products, we can easily make sure that the equality

$$(T\alpha z,\ u) = \frac{1}{4}\{(T(\alpha z + u),\ (\alpha z + u))$$
$$- (T(\alpha z - u),\ (\alpha z - u))\}$$

holds true. Using this equality and bearing in mind estimate (4), we get the relationship

$$\| Tz \|^2 \leqslant \frac{1}{4}\{|\, (T(\alpha z + u),\ (\alpha z + u))\, |$$
$$+ |\, (T(\alpha z - u),\ (\alpha z - u))\, |\}$$
$$\leqslant \{C_T (\| \alpha z + u \|^2 + \| \alpha z - u \|^2)\}.$$

Taking into consideration Property (4) of scalar products, written in the preceding section in terms of the norm, and the definitions of the quantities u and α, we obtain the inequality

$$\| Tz \|^2 \leqslant \frac{C_T}{4}(2\alpha^2 \| z \|^2 + 2 \| u \|^2)$$
$$= \frac{C_T}{2}\left(\frac{\| Tz \|}{\| z \|} \| z \|^2 + \frac{\| z \|}{\| Tz \|} \| Tz \|^2\right) = C_T \| z \| \| Tz \|.$$

Dividing both sides of the inequality by $\| Tz \|$, we derive the estimate $\| Tz \| \leqslant C_T \| z \|$, whence it follows that

$$\| T \|_1 \leqslant C_T. \tag{6}$$

The equality $C_T = \| T \|_1$ follows from estimates (5) and (6) and then the statement of the lemma follows from (4).

3. Realness of Characteristic Values (or Eigenvalues) of a Self-adjoint Operator. In a Hilbert space H, consider the equation

$$\varphi - \lambda T\varphi = 0, \tag{7}$$

where T is an operator mapping H into H, and λ is a constant number. It is clear that equation (7) has the trivial solution $\varphi = \theta$. It may turn out that for a certain value of $\lambda = \lambda^*$ equation (7) has a nontrivial solution $\varphi = \varphi^* \neq \neq \theta$. Then this value λ^* is called the *eigenvalue* (or *characteristic value*) *of the operator* T, and the solution φ^* corresponding to this λ^* is termed the *eigenfunction of the operator* T. The existence and properties of the eigenvalues and eigenfunctions of an arbitrary operator T is a rather complicated question, but for self-adjoint operators, the eigenfunctions and eigenvalues possess a number of interesting properties which we are going to prove.

Theorem 1. *Let T be a self-adjoint operator in a Hilbert space H. Then the eigenvalues of this operator are real.*

Let λ_0 be an eigenvalue, and φ_0 the appropriate eigenfunction of the operator T. This means that

$$\varphi_0 - \lambda_0 T\varphi_0 = 0, \tag{8}$$

where $\varphi_0 \neq \theta$, $\lambda_0 \neq 0$, and $\| \varphi_0 \|^2 \neq 0$. After scalar-multiplying (8) by φ_0, we arrive at the relationship

$$(\varphi_0, \varphi_0) = \lambda_0 (T\varphi_0, \varphi_0),$$

from which we obtain the representation

$$\frac{1}{\lambda_0} = \frac{(T\varphi_0, \varphi_0)}{(\varphi_0, \varphi_0)} = \frac{(T\varphi_0, \varphi_0)}{\| \varphi_0 \|^2}.$$

The statement of the theorem will be proved if we show that $(T\varphi_0, \varphi_0)$ is real. From the conditions of self-adjointness and properties of scalar products we have the equalities

$$(T\varphi_0, \varphi_0) = (\varphi_0, T\varphi_0) \text{ and } (T\varphi_0, \varphi_0) = \overline{(\varphi_0, T\varphi_0)},$$

from which we derive the relationship $(\varphi_0, T\varphi_0) = \overline{(\varphi_0, T\varphi_0)}$, which is possible only for real numbers. Consequently, $(T\varphi_0, \varphi_0)$ is real.

4. Orthogonality of Eigenfunctions of a Self-adjoint Operator. The eigenfuctions of self-adjoint operators possess a number of remarkable properties, orthogonality being one of them.

Theorem 2. *The eigenfunctions of a self-adjoint operator T associated with different eigenvalues are orthogonal.*

Let λ_1 and λ_2 be the eigenvalues of the operator T ($\lambda_1 \neq \lambda_2$), and φ_1 and φ_2 the eigenfunctions associated with them. This means that the following equalities are true:

$$\varphi_1 - \lambda_1 T\varphi_1 = 0 \text{ and } \varphi_2 - \lambda_2 T\varphi_2 = 0.$$

After scalar-multiplying the first equality by φ_2, we obtain the relationship

$$(\varphi_1, \varphi_2) = \lambda_1 (T\varphi_1, \varphi_2) = \lambda_1 (\varphi_1, T\varphi_2).$$

But $T\varphi_2 = (1/\lambda)\varphi_2$, and, since λ_2 is real, we come to the equality

$$(\varphi_1, \varphi_2) = (\lambda_1/\lambda_2)(\varphi_1, \varphi_2),$$

that is, to the equality $(1 - \lambda_1/\lambda_2)(\varphi_1, \varphi_2) = 0$. By the supposition, $\lambda_1 \neq \lambda_2$, hence the equality $(\varphi_1, \varphi_2) = 0$ follows.

Let us note without proof that the set of the eigenvalues of a completely continuous self-adjoint operator is at most countable. This assertion can be deduced from Theorem 6 to be proved later.

Theorem 3. *The sequence of eigenfunctions of a self-adjoint operator T can be made orthonormal.*

If to a certain eigenvalue there correspond several eigenfunctions (linearly independent), then they can be orthogonalized by applying Schmidt's orthogonalization method. And since the eigenfunctions corresponding to distinct eigenvalues are orthogonal, by norming them, we obtain an orthonormal system. Thus, we have a system of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$ and an associated with them orthonormal system of eigenfunctions $\varphi_1, \varphi_2, \ldots, \varphi_n, \ldots$ ($(\varphi_k, \varphi_m) = 0$ for $k \neq m$, $(\varphi_k, \varphi_k) = 1$), moreover equal λ_s are repeated as many times as the number of distinct eigenfunctions associated with them.

5. Existence of Eigenvalues. The previous theorems established the properties of eigenvalues and eigenfunctions

having assumed they exist. And the answer to the question concerning the existence of eigenvalues is given by the following theorem.

Theorem 4. *Any completely continuous self-adjoint operator T has at least one eigenvector φ^* associated with a nonzero eigenvalue λ^*.*

Let us denote the norm of the operator T by M. Then, according to Lemma 1, we can write the relationship

$$M = \| T \|_1 = \sup_{\|\varphi\|=1} | (T\varphi, \varphi) |,$$

from which we conclude that there exists a sequence of vectors $\{\varphi_n\}$, $\| \varphi_n \| = 1$ such that

$$\lim_{n\to\infty} (T\varphi_n, \varphi_n) = M \quad (\text{or} = -M).$$

The operator T is completely continuous, therefore it is possible to separate a convergent in H subsequence $\{\varphi_{n_k}\} \subset \subset \{\varphi_n\}$, for which there exists the limit

$$\lim_{k\to\infty} T\varphi_{n_k} = g \quad (\text{by norm } H).$$

Let us now prove the validity of the equality

$$\lim_{k\to\infty} \left\| \varphi_{n_k} - \frac{1}{M} T\varphi_{n_k} \right\| = 0. \tag{9}$$

Taking into account the relationship

$$\left\| \varphi_{n_k} - \frac{1}{M} T\varphi_{n_k} \right\|^2 = 1 - \frac{2}{M} (\varphi_{n_k}, T\varphi_{n_k}) + \frac{1}{M^2} \| T\varphi_{n_k} \|^2,$$

we arrive at the expression

$$\lim_{k\to\infty} \left\| \varphi_{n_k} - \frac{1}{M} T\varphi_{n_k} \right\|^2 = 1 - \frac{2}{M} \cdot M + \frac{1}{M} \| g \|^2$$

$$= \frac{\| g \|^2}{M} - 1 \geqslant 0. \tag{10}$$

Besides, from the definition of the norm there follows the estimate

$$\| T\varphi_{n_k} \| \leqslant \| T \|_1 \| \varphi_{n_k} \| = M,$$

whence, on passing to the limit, we get $\| g \| \leqslant M$.

Now taking into account inequality (10), we come to the equality $\| g \| = M$. Thus, relationship (9) holds true and it follows from this relationship that for the subsequence $\{\varphi_{n_k}\}$ in the metric of H there exists the limit $\lim_{k \to \infty} \varphi_{n_k} = -g/M$. Choosing $\varphi^* = g/M$ from relationship (9), we obtain

$$\varphi^* - \frac{1}{M} T\varphi^* = 0, \quad \text{or} \quad \varphi^* - \lambda^* T\varphi^* = 0,$$

where $\lambda^* = 1/M$.

6. Estimates of Growth of Eigenvalues. In applications of self-adjoint operators, especially in approximate solution of equations, we often encounter problems concerning the order of growth of eigenvalues. With this aim in mind, let us obtain some estimates for them.

Theorem 5. *For the eigenvalues λ_k, $k = 1, 2, \ldots$, of a completely continuous in H self-adjoint operator T the following estimate holds true:*

$$|\lambda_k| \geqslant \frac{1}{M}, \quad k = 1, 2, \ldots, \quad \text{where } M = \| T \|_1 = \sup_{\|\varphi\| = 1} |(T\varphi, \varphi)|.$$

Suppose the contrary, that is, let there exist an eigenvalue λ^* of the operator T such that $|\lambda^*| < 1/M$. Let φ^* be the eigenvector associated with λ^* and $\| \varphi^* \| = 1$ (otherwise we would consider the vector $\tilde{\varphi} = \varphi^*/\| \varphi^* \|$). After scalar-multiplying the equality $\varphi^* - \lambda^* T\varphi^* = 0$ by φ^*, we get the relationship

$$(\varphi^*, \varphi^*) = \lambda^* (T\varphi^*, \varphi^*) = |\lambda^*| |(T\varphi^*, \varphi^*)|,$$

from which there follows the inequality

$$\| \varphi^* \|^2 = 1 = |\lambda^*| |(T\varphi^*, \varphi^*)| \leqslant |\lambda^*| \sup_{\|\varphi\| = 1} |(T\varphi, \varphi)| = |\lambda^*| M < 1.$$

Thus, we have come to a contradiction; hence, $|\lambda^*| \geqslant 1/M$.

Remark. It follows from Theorem 5 that the eigenvalue $\lambda^* = 1/M = 1/\| T \|_1$ found in Theorem 4 is the least (by modulus) eigenvalue of the operator T, and therefore for

the norm of the operator T the following estimate holds true:

$$\| T \|_1 \leqslant \frac{1}{| \lambda_k |^2}, \tag{11}$$

where λ_k is the least (by modulus) eigenvalue of this operator.

In order to draw some conclusions concerning the behaviour of eigenvalues on the real axis, let us prove one auxiliary result.

Lemma 2. *Let T be a completely continuous operator mapping H into H, and let $\{g_k\}_{k=0}^{\infty}$ be an infinite orthonormal system of vectors in H. If the equalities*

$$Tg_k = \sum_{\nu=0}^{k} \beta_{k\nu} g_\nu, \quad k = 1, 2, \ldots,$$

are valid, then the relationship $\lim\limits_{k \to \infty} \beta_{kk} = 0$ *is fulfilled.*

If $n > m$, then we derive the inequality

$$\begin{aligned} \| Tg_n - Tg_m \|^2 &= \| \beta_{n,n} g_n + \ldots + \beta_{n,m+1} g_{m+1} \\ &\quad + (\beta_{n,m} - \beta_{m,m}) g_m + \ldots + (\beta_{n,0} - \beta_{m,0}) g_0 \|^2 \\ &= | \beta_{n,n} |^2 + \ldots + | \beta_{n,m+1} |^2 + | \beta_{n,m} - \beta_{m,m} |^2 \\ &\quad + \ldots + | \beta_{n,0} - \beta_{m,0} |^2 \geqslant | \beta_{n,n} |^2. \end{aligned}$$

Assuming that $\lim\limits_{k \to \infty} \beta_{kk} \neq 0$, we separate a subsequence of numbers $\| n_j \|$ for which the relationship $| \beta_{n_j, n_j} | \geqslant \geqslant \delta > 0$ is fulfilled. This means that

$$\| Tg_{n_k} - Tg_{n_m} \|^2 \geqslant \delta^2 \text{ for } n_k > n_m,$$

that is, it is impossible to separate a convergent subsequence from the infinite subsequence $\{Tg_{n_\nu}\}$, which contradicts the complete continuity of the operator T. The contradiction obtained proves the lemma.

Theorem 6. *Any completely continuous operator T: $(H \to H)$ for any $R > 0$ in the circle $| \lambda | < R$ can have only a finite number of eigenvalues.*

Suppose that for $| \lambda | < R$ there is an infinite number of eigenvalues $\lambda_1, \ldots, \lambda_n, \ldots, | \lambda_n | < R$. An infinite system of eigenvectors $\{\varphi_\nu\}$ corresponds to them, that is, we

have the equalities

$$\varphi_\nu - \lambda_\nu T\varphi_\nu = 0, \quad \nu = 1, 2, \ldots .$$

Orthogonalizing the vectors $\{\varphi_\nu\}$, we obtain an orthonormal system:

$$\begin{aligned} g_1 &= \alpha_{11}\varphi_1, \\ g_2 &= \alpha_{21}\varphi_1 + \alpha_{22}\varphi_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ g_k &= \alpha_{k1}\varphi_1 + \ldots + \alpha_{kk}\varphi_k, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{12}$$

Considering the relationships

$$Tg_k = \sum_{\nu=1}^{k} \alpha_{k\nu} T\varphi_\nu = \sum_{\nu=1}^{k} \frac{\alpha_{k\nu}}{\lambda_\nu} \varphi_\nu$$

find taking into account the reversibility of system (12), we and the representation

$$g_k - \lambda_k Tg_k = \sum_{\nu=1}^{k-1} \alpha_{k\nu} \left(1 - \frac{\lambda_k}{\lambda_\nu}\right) \varphi_\nu = -\sum_{\nu=1}^{k-1} \beta'_{k\nu}\varphi_\nu$$

$$= -\sum_{\nu=1}^{k-1} \beta_{k\nu} g_\nu .$$

Hence we see that the value Tg_k has the form

$$Tg_k = \sum_{\nu=1}^{k-1} \frac{\beta_{k\nu}}{\lambda_k} g_\nu + \frac{1}{\lambda_k} g_k,$$

where, by Lemma 2, $(1/\lambda_k) \to 0$ as $k \to \infty$. But this contradicts the fact that $|\lambda_k| < R$.

Sec. 5.3. HILBERT-SCHMIDT THEOREM AND ITS APPLICATION

1. Expanding the Values of an Operator into a Series. When solving operator equations in Hilbert spaces, we rather

frequently make use of the method of expanding the values of an operator into a series of eigenvectors of this operator. This expansion is characterized by the following theorem.

Theorem (Hilbert-Schmidt Theorem). *Let T be a completely continuous self-adjoint operator mapping H into H. Then for any $h \in H$ the vector $f = Th \in H$ can be represented in the form of a series in the eigenvectors of the operator T converging to f according to the metric of H, that is, if $\{\varphi_n\}_{n=1}^{\infty}$ is an orthonormal system of the eigenvectors of the operator T, then*

$$f = Th = \sum_{n=1}^{\infty} c_n \varphi_n \quad (\textit{in the metric of } H),$$

where

$$c_k = \frac{1}{\lambda_k}(h, \varphi_k), \quad k = 1, 2, \ldots .$$

Let $\varphi_1, \ldots, \varphi_n, \ldots$ be an orthonormal system of eigenvectors of the operator T corresponding to the sequence $\{\lambda_n\}$ of eigenvalues of this operator arranged in increasing order of moduli $|\lambda_1| \leqslant |\lambda_2| \leqslant \ldots \leqslant |\lambda_n| \leqslant \ldots$. The vector $h \in H$ is associated with the series with respect to the system $\{\varphi_n\}$, that is,

$$h \sim \sum_{n=1}^{\infty} h_n \varphi_n, \quad \text{where} \quad h_n = (h, \varphi_n).$$

Applying Bessel's inequality, we conclude that the series $\sum_{n=1}^{\infty} |h_n|^2$ converges. Let us also form a series for the vector $f = Th$:

$$f \sim \sum_{n=1}^{\infty} f_n \varphi_n, \quad \text{where} \quad f_n = (f, \varphi_n).$$

Taking into consideration that the operator T is self-adjoint and bearing in mind the equality $\varphi_n - \lambda_n T\varphi_n = 0$, we find the coefficients f_n:

$$f_n = (Th, \varphi_n) = (h, T\varphi_n) = \left(h, \frac{1}{\lambda_n}\varphi_n\right) = \frac{h_n}{\lambda_n}.$$

Consequently, the series for the vector $f = Th$ takes the form

$$f \sim \sum_{n=1}^{\infty} \frac{h_n}{\lambda_n} \varphi_n = \sum_{n=1}^{\infty} \frac{(h, \varphi_n)}{\lambda_n} \varphi_n.$$

Let

$$S_n h = \sum_{k=1}^{n} \frac{(h, \varphi_k)}{\lambda_k} \varphi_k, \quad n = 1, 2, \ldots,$$

be partial sums for $f = Th$. Let us introduce the operator

$$T^{(n)}h = f - S_n h = Th - S_n h$$

and prove that the eigenvalues of the operator $T^{(n)}$ and associated with them eigenvectors are contained among those eigenvalues and eigenvectors of the operator T whose ordinal numbers exceed n.

Really, if $\varphi^* - \mu T^{(n)}\varphi^* = 0$, then, substituting the value of the operator $T^{(n)}$, we obtain the equality

$$\varphi^* - \mu \left(T\varphi^* - \sum_{k=1}^{n} \frac{(\varphi^*, \varphi_k)}{\lambda_k} \varphi_k \right) = 0. \tag{1}$$

After scalar-multiplying this equality by φ_l for any $l \leqslant n$ and bearing in mind that the operator T is self-adjoint, we arrive at the expression

$$(\varphi^*, \varphi_l) - \mu (T\varphi^*, \varphi_l) + \mu \sum_{k=1}^{n} \frac{(\varphi^*, \varphi_k)}{\lambda_k} (\varphi_k, \varphi_l) = 0,$$

from which, by virtue of the orthonormality of the system $\{\varphi_n\}$, we have the relationship

$$(\varphi^*, \varphi_l) - \mu (\varphi^*, T\varphi_l) + \frac{\mu}{\lambda_l} (\varphi^*, \varphi_l) = 0.$$

Taking into consideration the equality

$$\frac{\mu}{\lambda_l} (\varphi^*, \varphi_l) = \mu (\varphi^*, T\varphi_l),$$

we obtain that $(\varphi^*, \varphi_l) = 0$ for all $l \leqslant n$.

Taking into account these equalities, we derive from (1) the relationship

$$\varphi^* - \mu T\varphi^* = 0.$$

Hence we conclude that μ is an eigenvalue of the operator T, and φ^* is the eigenvector corresponding to this eigenvalue which is orthogonal to the first n vectors of the system $\{\varphi_k\}$, that is, the ordinal number of the vector φ^* in this system exceeds n.

Let us now assume that the sequence $\{\varphi_k\}$ consists of a finite number of vectors $\varphi_1, \varphi_2, \ldots, \varphi_m$. Then the self-adjoint operator $T^{(n)}$ for $n > m$ has no eigenvectors orthogonal to all the vectors of this system, that is, it has not a single nonzero eigenvalue. This means that $T^{(n)}h = 0$ for any $h \in H$ and any $n > m$; we represent the vector $f = Th$ as a finite "polynomial" with respect to the system $\varphi_1, \varphi_2, \ldots, \varphi_m$:

$$f = Th = \sum_{k=1}^{m} \frac{(h, \varphi_k)}{\lambda_k} \varphi_k.$$

And if the sequence $\{\varphi_k\}$ is infinite, then, by virtue of Theorem 6 proved in the preceding section, $\lim\limits_{k\to\infty} \lambda_k = \infty$. Let us show that λ_{n+1} is the least (by modulus) eigenvalue of the operator $T^{(n)}$. The equalities

$$\varphi_{n+1} - \lambda_{n+1}T^{(n)}\varphi_{n+1} = \varphi_{n+1} - \lambda_{n+1}T\varphi_{n+1}$$

$$- \lambda_{n+1} \sum_{k=1}^{n} \frac{(\varphi_k, \varphi_{n+1})}{\lambda_k} \varphi_k = \varphi_{n+1} - \lambda_{n+1}T\varphi_{n+1} = 0$$

hold true, and since, by the hypothesis, $|\lambda_1| \leqslant |\lambda_2| \leqslant \ldots \leqslant |\lambda_n| \leqslant \ldots$, it follows that λ_{n+1} is the least (by modulus) eigenvalue of the operator $T^{(n)}$. Applying inequality (11) from Sec. 5.2, we get the estimate

$$\| T^{(n)} \|_1 \leqslant \frac{1}{|\lambda_{n+1}|}.$$

Thus, for any $h \in H$ there holds the inequality

$$\| T^{(n)}h \| \leqslant \| T^{(n)} \|_1 \| h \| \leqslant \frac{1}{|\lambda_{n+1}|} \| h \|,$$

from which there follows the relationship

$$\lim_{n\to\infty} \| T^{(n)}h \| \leqslant \| h \| \lim_{n\to\infty} \frac{1}{|\lambda_{n+1}|} = 0,$$

and this means that

$$\| Th - S_n h \| \to 0 \text{ as } n \to \infty.$$

2. Solving Symmetric Integral Equations. Let us apply the Hilbert-Schmidt theorem proved above to solve the following problem.

Find the solution $\varphi(x) \in L^2[a, b]$ *of the Fredholm symmetric integral equation*

$$\varphi(x) - \lambda \int_a^b K(x, s)\varphi(s)\,ds = f(x), \tag{2}$$

assuming that the system of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$ *of the operator* $T\varphi = \int_a^b K(x, s)\varphi(s)\,ds$ *and the system of its eigenfunctions* $\varphi_1(x), \ldots, \varphi_n(x), \ldots$ *are known.*

We will consider separately the case of a regular λ and the case when λ is an eigenvalue.

Let λ be a regular number. We find the solution $\varphi(x) \in L^2[a, b]$. Applying the Hilbert-Schmidt theorem, we may write the equality

$$\int_a^b K(x, s)\varphi(s)\,ds = \sum_{n=1}^{\infty} \frac{(\varphi, \varphi_n)}{\lambda_n}\varphi_n(x).$$

Substituting this expression into equation (2), we get the relationship

$$\varphi(x) - \lambda \sum_{n=1}^{\infty} \frac{(\varphi, \varphi_n)}{\lambda_n}\varphi_n(x) = f(x). \tag{3}$$

After scalar-multiplying equality (3) by $\varphi_m(x)$, we come to the relationships

$$(\varphi, \varphi_m) - \frac{\lambda}{\lambda_m}(\varphi, \varphi_m)(\varphi_m, \varphi_m) = (f, \varphi_m), \quad m = 1, 2, \ldots, \tag{4}$$

and since λ is a regular number, hence we find the coefficients $(\varphi, \varphi_m) = c_m$:

$$c_m = \frac{\lambda_m (f, \varphi_m)}{\lambda_m - \lambda} = \frac{\lambda_m f_m}{\lambda_m - \lambda}.$$

Substituting these coefficients into relationship (3), we get the desired solution:

$$\varphi(x) = \lambda \sum_{n=1}^{\infty} \frac{f_n}{\lambda_n - \lambda} \varphi_n(x) + f(x). \tag{5}$$

Let now λ be an eigenvalue, i.e.

$$\lambda = \lambda_p = \lambda_{p+1} = \ldots = \lambda_q, \quad q \geqslant p.$$

Then from expressions (4) it is possible to determine c_m only for $m \neq p, p+1, \ldots, q$, and, consequently, we have

$$c_m = \frac{\lambda_m f_m}{\lambda_m - \lambda} \quad \text{if} \quad m \neq p, \; p+1, \ldots, q.$$

But if the number m is such that $p \leqslant m \leqslant q$, then equalities (4) take the form

$$c_m \left(1 - \frac{\lambda}{\lambda_m}\right) = (f, \varphi_m), \text{ where } 1 - \frac{\lambda}{\lambda_m} = 0.$$

These equalities are possible only if $(f, \varphi_m) = 0$. Therefore for solvability of equation (2) it is necessary that the constant term of $f(x)$ be orthogonal to all the eigenfunctions associated with the given eigenvalue λ. Thus, if $f(x)$ satisfies the conditions

$$(f, \varphi_m) = 0, \quad m = p, \quad p+1, \ldots, q, \tag{6}$$

then equation (2) has infinitely many solutions determined by the formula

$$\varphi(x) = f(x) + \sum_{n=1}^{\infty} \frac{\lambda_n}{\lambda_n - \lambda} \varphi_n(x) + \sum_{n=p}^{q} d_n \varphi_n(x),$$

where d_n are arbitrary constants, and the prime in the sum means that the dummy indices $n = p, p+1, \ldots, q$ are omitted.

And if the function $f(x)$ does not satisfy conditions (6), and $\lambda = \lambda_p = \ldots = \lambda_q$, then equation (2) has no solution.

Let us now consider a few examples illustrating the foregoing.

Example 1°. Find the solution of the equation

$$\varphi(x) - \lambda \int_0^{\pi/2} K(x, t)\, \varphi(t)\, dt = 1 \tag{7}$$

if

$$K(x, t) = \begin{cases} \sin x \cos t & \text{for} \quad 0 \leqslant x \leqslant t \leqslant \pi/2, \\ \sin t \cos x & \text{for} \quad 0 \leqslant t \leqslant x \leqslant \pi/2. \end{cases} \tag{8}$$

First of all we find the eigenvalues and eigenfunctions of the equation

$$\varphi(x) - \lambda \int_0^{\pi/2} K(x, t)\, \varphi(t)\, dt = 0,$$

which we represent in the form

$$\varphi(x) - \lambda \int_0^x \sin t \cos x\, \varphi(t)\, dt - \lambda \int_x^{\pi/2} \sin x \cos t\, \varphi(t)\, dt = 0.$$

We then differentiate this equality twice:

$$\varphi'(x) + \lambda \sin x \int_0^x \sin t\, \varphi(t)\, dt - \lambda \sin x \cos x\, \varphi(x) - \lambda \cos x \int_x^{\pi/2} \cos t\, \varphi(t)\, dt + \lambda \sin x \cos x\, \varphi(x) = 0,$$

i.e.

$$\varphi'(x) + \lambda \sin x \int_0^x \sin t\, \varphi(t)\, dt - \lambda \cos x \int_x^{\pi/2} \cos t\, \varphi(t)\, dt = 0;$$

$$\varphi''(x)+\lambda\cos x\int_0^x \sin t\,\varphi(t)\,dt$$
$$+\lambda\sin x\sin x\,\varphi(x)+\lambda\sin x\int_x^{\pi/2}\cos t\,\varphi(t)\,dt$$
$$+\lambda\cos x\cos x\,\varphi(x)=0.$$

Hence, taking into account representation (8), we obtain the equation

$$\varphi''(x)+\varphi(x)+\lambda\varphi(x)=0 \tag{9}$$

and the boundary conditions

$$\varphi(0)=\varphi(\pi/2)=0. \tag{10}$$

Thus, equation (7) is reduced to equation (9) with the boundary conditions (10), that is, to the Sturm-Liouville problem. Let us find the solution of this problem.

(a) Let $\lambda+1<0$, i.e. $\lambda<-1$. Then the general solution of equation (9) is the function

$$\varphi(x)=C_1e^{\sqrt{-\lambda-1}x}+C_2e^{-\sqrt{-\lambda-1}x},$$

and the boundary conditions (10) lead to the system of equations

$$C_1+C_2=0,$$
$$C_1(e^{\sqrt{-\lambda-1}(\pi/2)}-e^{-\sqrt{-\lambda-1}(\pi/2)})=0,$$

whose solutions are the numbers $C_1=C_2=0$. This means that all the values $\lambda<-1$ are regular.

(b) Let $\lambda+1=0$, i.e. $\lambda=-1$. The general solution of equation (9) is the function

$$\varphi(x)=C_1x+C_2,$$

and from the boundary conditions (10) we once again derive the relationships

$$C_2=0 \text{ and } C_1(\pi/2)=0, \text{ i.e. } C_1=C_2=0.$$

Thus, $\lambda=-1$ is a regular value of the parameter λ,

(c) Let now $\lambda + 1 > 0$, i.e. $\lambda > -1$. Then the solution of equation (9) has the form

$$\varphi(x) = C_1 \cos \sqrt{\lambda + 1}\, x + C_2 \sin \sqrt{\lambda + 1}\, x.$$

Taking into account the boundary conditions (10), we get the relationships

$$C_1 = 0 \text{ and } \varphi(\pi/2) = C_2 \sin \sqrt{\lambda + 1}\, (\pi/2) = 0.$$

The solution will not be trivial only in the case when $\sin \sqrt{\lambda + 1}\, (\pi/2) = 0$, that is, if $\sqrt{\lambda + 1}\, (\pi/2) = k\pi$. This means that if $\lambda = \lambda_k = 4k^2 - 1$, $k = 1, 2, \ldots$, then equation (9) has nontrivial solutions

$$\varphi_k(x) = \sin 2kx, \quad k = 1, 2, \ldots.$$

Bearing this in mind, we may pass over to solving equation (7). In this case $f(x) = 1$ and

$$f_k = (f, \varphi_k) = \frac{4}{\pi} \int_0^{\pi/2} 1 \cdot \sin 2kx \, dx = -\frac{4}{\pi} \frac{\cos 2kx}{2k} \Big|_0^{\pi/2}$$

$$= -\frac{2}{\pi} (\cos k\pi - 1)$$

$$= \begin{cases} 0 & \text{for } k = 2m, \\ \dfrac{4}{\pi(2m-1)} & \text{for } k = 2m-1, \end{cases} \quad m = 1, 2, \ldots,$$

therefore, applying formula (5) for $\lambda \neq \lambda_k$, we find the desired solution:

$$\varphi(x) = 1 + \lambda \sum_{m=1}^{\infty} \frac{4 \sin (4m-2)\, x}{\pi (2m-1) \{[4(2m-1)^2 - 1] - \lambda\}}$$

$$= 1 + \frac{4\lambda}{\pi} \sum_{m=1}^{\infty} \frac{\sin (4m-2)\, x}{(2m-1)(16m^2 - 16m + 3 - \lambda)}. \qquad (11)$$

And if $\lambda = \lambda_{2\nu} = 16\nu^2 - 1$, then $f_{2\nu} = (f, \varphi_{2\nu}) = 0$, i.e. the function $f(x) \equiv 1$ is orthogonal to the corresponding eigenfunction $\varphi_{2\nu}(x) = \sin 4\nu x$, and therefore the solution of equation (7) will differ from solution (11) by the additional term $C \sin 4\nu x$, where C is an arbitrary constant.

If $\lambda = \lambda_{2\nu-1} = 16\nu^2 - 16\nu + 3$, $\nu = 1, 2, \ldots$, equation (7) has no solution.

Example 2°. Solve the equation

$$\varphi(x) - \lambda \int_0^1 e^{-|x-t|} \varphi(t)\, dt = x. \tag{12}$$

We write the homogeneous equation in the form

$$\varphi(x) - \lambda e^{-x} \int_0^x e^t \varphi(t)\, dt - \lambda e^x \int_x^1 e^{-t} \varphi(t)\, dt = 0. \tag{13}$$

Differentiating this equality twice, we obtain the relationships

$$\varphi'(x) + \lambda e^{-x} \int_0^x e^t \varphi(t)\, dt - \lambda e^x \int_x^1 e^{-t} \varphi(t)\, dt = 0 \tag{14}$$

and

$$\varphi''(x) - \lambda e^{-x} \int_0^x e^t \varphi(t)\, dt - \lambda e^x \int_x^1 e^{-t} \varphi(t)\, dt + \lambda\varphi(x) + \lambda\varphi(x) = 0.$$

By virtue of (13), the last relationship is written in the form

$$\varphi''(x) + (2\lambda - 1)\varphi(x) = 0. \tag{15}$$

From equalities (13) and (14) there follow the boundary conditions

$$\varphi(0) = \lambda \int_0^1 e^{-t} \varphi(t)\, dt = \varphi'(0),$$

$$\varphi(1) = \frac{\lambda}{e} \int_0^1 e^t \varphi(t)\, dt = -\varphi'(1),$$

that is, the conditions

$$\varphi(0) = \varphi'(0), \quad \varphi(1) = -\varphi'(1). \tag{16}$$

If $2\lambda - 1 < 0$, i.e. $\lambda < 1/2$, then for the solution $\varphi(x)$ of equation (15) we have the relationships

$$\varphi(x) = C_1 e^{\sqrt{1-2\lambda}x} + C_2 e^{-\sqrt{1-2\lambda}x}$$

and

$$\varphi'(x) = \sqrt{1-2\lambda}\,(C_1 e^{\sqrt{1-2\lambda}x} - C_2 e^{-\sqrt{1-2\lambda}x}).$$

Conditions (16) then lead to the system of equations

$$C_1 + C_2 = \sqrt{1-2\lambda}\,(C_1 - C_2),$$
$$C_1 e^{\sqrt{1-2\lambda}} + C_2 e^{-\sqrt{1-2\lambda}} = -\sqrt{1-2\lambda}\,(C_1 e^{\sqrt{1-2\lambda}} - C_2 e^{-\sqrt{1-2\lambda}})$$

which, on introducing the notation $\sqrt{1-2\lambda} = \mu$, $\mu > 0$, will be written in the form

$$\begin{aligned} C_1(1-\mu) + C_2(1+\mu) &= 0, \\ C_1 e^{\mu}(1+\mu) + C_2 e^{-\mu}(1-\mu) &= 0. \end{aligned} \tag{17}$$

The determinant of this system, i.e. the expression

$$e^{-\mu}(1-\mu)^2 - e^{\mu}(1+\mu)^2 = e^{-\mu}(1+\mu)^2\left[\left(\frac{1-\mu}{1+\mu}\right)^2 - e^{2\mu}\right],$$

may vanish only if

$$e^{\mu} = \frac{1-\mu}{1+\mu}. \tag{18}$$

But for $\mu > 0$ the right-hand member is less than unity, while the left-hand member exceeds unity, that is, equality (18) is impossible, and this means that system (17) has only a trivial solution: $C_1 = C_2 = 0$, i.e. the values $\lambda < 1/2$ are regular.

And if $\lambda = 1/2$, then for solving equation (15) we have the equalities

$$\varphi(x) = C_1 x + C_2, \quad \varphi'(x) = C_1.$$

Whence, taking into account (16), we obtain the relationships

$$C_2 = C_1, \quad C_1 + C_2 = C_1, \quad \text{i.e. } C_2 = C_1 = 0.$$

Finally, for $\lambda > 1/2$ the solution of equation (15) will be the function

$$\varphi(x) = C_1 \cos\sqrt{2\lambda - 1}\,x + C_2 \sin\sqrt{2\lambda - 1}\,x.$$

Setting once again $\mu = \sqrt{2\lambda - 1}$, we write the equalities

$$\varphi(x) = C_1 \cos \mu x + C_2 \sin \mu x,$$
$$\varphi'(x) = \mu(-C_1 \sin \mu x + C_2 \cos \mu x)$$

and, taking into account conditions (16), we come to the system of equations

$$C_1 = \mu C_2,$$
$$C_1 \cos \mu + C_2 \sin \mu = -\mu(-C_1 \sin \mu + C_2 \cos \mu).$$

Thus, for C_1 and C_2 we have the following homogeneous system of equations:

$$\begin{aligned} C_1 - \mu C_2 &= 0, \\ C_1(\cos \mu - \mu \sin \mu) + C_2(\sin \mu + \mu \cos \mu) &= 0. \end{aligned} \tag{19}$$

Equating the determinant of this system to zero, we get

$$\sin \mu + 2\mu \cos \mu - \mu^2 \sin \mu = 0,$$

or

$$(1 - \mu^2) \sin \mu + 2\mu \cos \mu = 0.$$

Consequently, system (19) has nonzero solutions if the quantity μ satisfies the transcendental equation

$$2 \cot \mu = \mu - \frac{1}{\mu}. \tag{20}$$

Denoting by μ_n the positive roots of equation (20) and choosing $C_2 = 1$, we find $C_1^{(n)} = \mu_n$, $n = 1, 2, \ldots$, and therefore the eigenfunctions of equation (15) will be written in the form

$$\varphi_n(x) = \mu_n \cos \mu_n x + \sin \mu_n x, \quad n = 1, 2, \ldots.$$

We introduce the notation

$$k_n^2 = \int_0^1 \varphi_n^2(x)\, dx, \quad n = 1, 2, \ldots, \tag{21}$$

and set

$$\psi_n(x) = \frac{1}{k_n} \varphi_n(x), \quad n = 1, 2, \ldots,$$

Then for $\lambda \neq \lambda_n = (1 + \mu_n^2)/2$ the solution of equation (12) will be written in the form

$$\varphi(x) = x + \lambda \sum_{n=1}^{\infty} \frac{(x, \psi_n)}{\lambda_n - \lambda} \psi_n(x)$$

$$= x + \lambda \sum_{n=1}^{\infty} \int_0^1 x\varphi_n(x)\, dx \frac{1}{k_n^2 (\lambda_n - \lambda)} \varphi_n(x).$$

Compute the integrals $\int_0^1 x\varphi_n(x)\, dx$:

$$\int_0^1 x\varphi_n(x)\, dx = \int_0^1 x(\mu_n \cos \mu_n x + \sin \mu_n x)\, dx$$

$$= \left[x \left(\sin \mu_n x - \frac{\cos \mu_n x}{\mu_n} \right) \right] \Big|_{x=0}^{1}$$

$$- \int_0^1 \left(\sin \mu_n x - \frac{\cos \mu_n x}{\mu_n} \right) dx = \sin \mu_n - \frac{\cos \mu_n}{\mu_n}$$

$$+ \frac{\cos \mu_n - 1}{\mu_n} + \frac{\sin \mu_n}{\mu_n^2} = \frac{(\mu_n^2 + 1) \sin \mu_n - \mu_n}{\mu_n^2}.$$

Substituting these values into the found solution, we obtain the expression

$$\varphi(x) = x + 2\lambda \sum_{n=1}^{\infty} \frac{(\mu_n^2 + 1) \sin \mu_n - \mu_n}{k_n^2 \mu_n^2 (\mu_n^2 + 1 - 2\lambda)} \varphi_n(x),$$

where the numbers k_n are determined by equalities (21).

CHAPTER 6

Fundamentals of the Calculus of Variations

Sec. 6.1. BASIC NOTIONS

1. Examples of Variation Problems. The calculus of variations investigates methods that permit finding the maximal and minimal values of functionals. Problems in which it is required to investigate a function for a maximum or a minimum are called *variation problems.*

Functionals are variable quantities whose values are determined by the choice of one or several functions. Thus, the functional is a generalization of the notion of function. The function $y = f(x)$ associates each value of x from a set of numbers X with a definite number y from a set of numbers Y. If now, instead of the set of numbers X, we consider the set R of an arbitrary nature (say, the set of functions), then we shall come to the notion of the functional. The *functional* $I(f)$ is said to be defined on the set R if every element $f \in R$ is associated with a certain number $I(f)$.

Let us give some examples of variation problems.

1. Finding the plane line of minimum length connecting two given points. In this problem the functional under investigation is the length of the line. The required line is the line segment joining the given points.

2. Finding the curve, connecting two given points, along which a material point slides down under the force of gravity in the shortest time. Here, the functional under investigation is the time during which the point moves along the curve. The curve minimizing this functional is called the *brachistochrone**.

* In 1696 Johann Bernoulli published a letter in which he advanced the problem of the line of quickest descent (brachistochrone). The problem of the brachistochrone was solved by Johann Bernoulli, Jacob Bernoulli, Newton, and L'Hospital.

3. Finding a closed curve of given length bounding a maximum area. Here, the functional under investigation is the area of a figure. As distinct from the two previous problems, in this problem the curves under consideration are imposed with the auxiliary peculiar condition that all the curves must have one and the same length, or in other words, the length of the curve must be constant. This additional condition is specified by a functional having a given value. Conditions of this kind are called *isoperimetric*. General methods for solving problems with isoperimetric conditions were elaborated by L. Euler.

2. Spaces of Continuously Differentiable Functions. We shall deal with variation problems for functionals specified in the spaces C, $C^{(1)}$, and $C^{(n)}$.

Let us introduce these spaces. The space C consists of functions $f(x)$ continuous on the interval $[a, b]$ and the norm is defined by the equality

$$\| f \|_C = \max_{x \in [a, b]} | f(x) |. \tag{1}$$

$C^{(1)}$ will denote the space of continuously differentiable on $[a, b]$ functions $f(x)$ with the norm

$$\| f \|_1 = \| f \|_{C^{(1)}} = \max_{x \in [a, b]} | f(x) | + \max_{x \in [a, b]} | f'(x) | = \| f \|_C + \| f' \|_C. \tag{2}$$

The space $C^{(n)}$ consists of functions $f(x)$ having on $[a, b]$ continuous derivatives $f^{(k)}(x)$, $k = 1, 2, \ldots, n$, up to the nth order inclusively, and the norm is defined by the equality

$$\| f \|_n = \| f \|_{C^{(n)}} = \sum_{k=0}^{n} \max_{x \in [a, b]} | f^{(k)}(x) | = \sum_{k=0}^{n} \| f^{(k)} \|_C, \tag{3}$$

where $f^{(0)}(x) = f(x)$.

By an ε-neighbourhood of the element y_0 of the normed space E we understand the set of all elements or points $y \in E$ such that $\rho(y_0, y) = \| y - y_0 \| < \varepsilon$. Consider the set R of functions having on $[a, b]$ continuous derivatives of up to the nth order inclusively. In this set, we can introduce the norm in both senses: as (1) and (2) or (3).

In the first case (for norm (1)), enclosed in the ε-neighbourhood of the curve $y = f(x)$ are those curves whose ordinates differ by less than ε. Thus, all the curves from R lying entirely between the curves $y = f(x) - \varepsilon$ and $y = f(x) + \varepsilon$ belong to the ε-neighbourhood of the curve $y = f(x)$ (Fig. 38).

But the closeness of the curves with respect to their ordinates does not mean that they are also close in the sense of

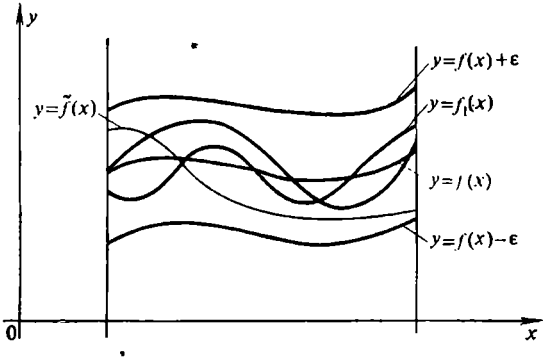


Fig. 38

norm (2). With respect to norm (2), the ε-neighbourhood of the curve $y = f(x)$ will contain only those curves which differ, but little, both in values of the function and those of the first derivative. For example, in Fig. 38 the curves $y = f(x)$ and $y = f_1(x)$ are close in the sense of the space C (norm (1)), but are not close in the sense of the space $C^{(1)}$ (norm (2)). If the curves $y = f(x)$ and $y = \tilde{f}(x)$ are close in the sense of the norm of the space $C^{(1)}$, then they are also close in the sense of the norm of the space C.

The closeness of elements in the sense of the space $C^{(n)}$ means the closeness of the values of both the functions themselves and their derivatives up to the nth order inclusively.

Let us recall the notion of continuity of a functional. If in the set R the notion of the distance $\rho(y, z)$, where $y, z \in R$ is defined, then the definition of continuity of a functional is introduced similarly to the definition of con-

tinuity of a function, namely: the functional $I(y)$ is said to be *continuous at the point* $y_0 \in R$ if $\forall \varepsilon > 0 \; \exists \delta > 0$ such that $|I(y_0) - I(y)| < \varepsilon$ as soon as $\rho(y_0, y) < \delta$. Hence it follows that the continuity of a functional depends not only on the analytic expression of this functional but also on the normed space in which it is specified. One and the same functional can be continuous in one normed space and discontinuous in the other.

Example 1°. Show that the functional

$$I(y) = \int_a^b \sqrt{1+y'^2}\, dx \tag{4}$$

defined on the set of functions $y = y(x)$ continuous together with the first derivative on the closed interval $[a, b]$ is continuous in the space $C^{(1)}$, but is discontinuous in C.

Let us first show that the given functional is continuous in the space $C^{(1)}$. Indeed, we have the relationships

$$|I(y) - I(y_0)| = \left| \int_a^b \left(\sqrt{1+y'^2} - \sqrt{1+y_0'^2}\right) dx \right|$$

$$\leqslant \int_a^b \frac{|y'+y_0'|\,|y'-y_0'|}{\sqrt{1+y'^2}+\sqrt{1+y_0'^2}}\, dx.$$

The continuity of the derivatives y' and y_0' implies that

$$\frac{|y'+y_0'|}{\sqrt{1+y'^2}+\sqrt{1+y_0'^2}} \leqslant M.$$

Further, from the definition of norm (2) in the space $C^{(1)}$ we conclude that

$$\max_{x \in [a,\, b]} |y' - y_0'| = \|y' - y_0'\|_C \leqslant \|y - y_0\|_1.$$

Given $\varepsilon > 0$, we choose $\delta = \varepsilon/(M(b-a))$. Then for all $y \in C^{(1)}$ and such that $\|y - y_0\|_1 < \delta$, we have

$$|I(y) - I(y_0)| \leqslant M \max_{x \in [a,\, b]} |y' - y_0'|\,(b-a)$$

$$\leqslant M(b-a)\,\|y - y_0\|_1 < \varepsilon.$$

and this means that the given functional is continuous in the space $C^{(1)}$.

When considered in the space C, the given functional is no longer continuous, since the norm $\| y - y_0 \|_C$ allows only for the closeness with respect to ordinates, but disregards the closeness of the tangents (see Fig. 38). Since it is impossible to estimate $\max | y' - y_0' |$ from $\max | y - y_0 |$, $\max | y' - y_0' |$ may exceed δ, whereas $\max | y - y_0 | < \delta$, that is, y will belong to the ε-neighbourhood of y_0 in the space C, but $| I(y) - I(y_0) | > \varepsilon$. And this means that the functional $I(y)$ will be discontinuous in the space C.

3. Linear Functional. Let us give the definition of a linear functional. Let E be a normed linear space. The functional $L(y)$ is called *linear* if: (a) it is continuous in the space E; (b) for any $y_1, y_2 \in E$ the additivity condition is fulfilled

$$L(y_1 + y_2) = L(y_1) + L(y_2). \tag{5}$$

It is possible to prove that the additivity condition and continuity of a linear functional imply its homogeneity, that is, if λ is an arbitrary real number, then

$$L(\lambda y) = \lambda L(y). \tag{6}$$

As an example of a linear functional, we can take the integral $L(f) = \int_a^b f(x)\,dx$ defined in the space C. It is readily verified that it is continuous and additive.

Example 2°. Show that the functional

$$L(f) = \int_a^b \alpha(x) f(x)\,dx, \tag{7}$$

where $\alpha(x)$ is a fixed continuous function, is linear in the space C.

The additivity of this functional is obvious. Let us show its continuity. Taking into consideration that $\alpha(x)$ is bounded ($| \alpha(x) | < M$), let us estimate the modulus of the difference; we have

$$| L(f) - L(f_1) | \leqslant \int_a^b | \alpha(x) | \, | f(x) - f_1(x) | \, dx$$

$$\leqslant M \max_{x \in [a, b]} | f(x) - f_1(x) | (b-a) \leqslant M \| f - f_1 \|_C (b-a) < \varepsilon,$$

as soon as the norm $\| f - f_1 \|_C < \varepsilon/(M(b-a))$. And this means that the given functional is continuous.

Functional (4) $I(y) = \int_a^b \sqrt{1 + y'^2}\, dx$ is not linear in the space C, since the conditions of its continuity and additivity are not fulfilled. Nor will it be linear in the space $C^{(1)}$; although it is continuous, but it is not additive.

Sec. 6.2. EXTREMUM OF A FUNCTIONAL

1. Variation of a Functional. The methods of solving variation problems, that is, problems involving the investigation of functionals for maxima and minima, are extremely similar to the methods of investigating functions for the least and greatest values.

Consider the functional $I(y)$ in the normed space E. The *increment*, or *variation*, δy *of the argument* y *of the functional* $I(y)$ is defined as the difference between two elements, $y, \tilde{y} \in E$, that is, $\delta y = \tilde{y} - y$. In variation problems, the variation δy plays the role analogous to the role of the increment of the independent variable Δx, and the variation of a functional the role of the differential of a function in problems involving the investigation of functions $f(x)$ for extrema.

The quantity $\Delta I = \Delta I(\delta y) = I(y + \delta y) - I(y)$ is called the *increment of a functional* corresponding to the increment δy. With y fixed, the increment $\Delta I(\delta y)$ represents the functional of δy. Suppose that the increment can be represented in the form of the sum

$$\Delta I = \Delta I(\delta y) = L(\delta y) + \varepsilon(\| \delta y \|), \qquad (1)$$

where $L(\delta y)$ is a functional linear with respect to the variation δy, and $\varepsilon(\| \delta y \|)$ is an infinitesimal of a higher order as compared with $\| \delta y \|$. The principal part $L(\delta y)$ of this increment ΔI representing the linear functional with respect to δy is called the *variation of the functional* $I(y)$ and is denoted by δI, that is, $\delta I = L(\delta y)$.

We are interested in the functionals $I(y)$ defined on a certain set of n-fold differentiable functions $y(x)$. In this case the variation of the argument $\delta y = \tilde{y}(x) - y(x)$ is a function of x which can be differentiated n times:

$$(\delta y)' = \tilde{y}' - y' = \delta y',$$
$$(\delta y)'' = \tilde{y}'' - y'' = \delta y'',$$
$$\dots\dots\dots\dots\dots$$
$$(\delta y)^{(n)} = \tilde{y}^{(n)} - y^{(n)} = \delta y^{(n)}.$$

These equalities mean that the derivatives of the variation of the function $y(x)$ are equal to the variation of the derivatives. Thus, if the function $y(x)$ receives the increment δy, then its first derivative $y'(x)$ receives the increment $(\delta y)' = \delta y'$, and the kth derivative $y^{(k)}(x)$, the increment $(\delta y)^{(k)} = \delta y^{(k)}$, $k = 1, 2, \dots, n$.

The calculus of variations deals with functionals of the form

$$I(y) = \int_a^b F(x, y, y')\,dx, \tag{2}$$

where $F(x, y, z)$ is a continuous function having continuous partial derivatives with respect to all variables up to the second order inclusively. It is possible to show that functional (2) is continuous in $C^{(1)}$, provided that the conditions imposed on $F(x, y, z)$ are fulfilled. Let us find the variation of this functional in the space $C^{(1)}$. Let $\delta y = h(x)$ be the increment of the function $y(x)$, then $\delta y' = h'$. Applying Taylor's formula to the difference of the integrands, we obtain for the increment ΔI the relationship

$$\Delta I = \int_a^b [F(x, y+h, y'+h') - F(x, y, y')]\,dx$$
$$= \int_a^b [F_y(x, y, y')\,h + F_{y'}(x, y, y')\,h']\,dx$$

$$+\frac{1}{2}\int_a^b [F_{yy}(x, y^*, y^{*\prime})\,h^2+2F_{yy'}(x, y^*, y^{*\prime})\,\widetilde{h}'$$
$$+F_{y'y'}(x, y^*, y^{*\prime})\,h'^2]\,dx,$$

where $y^* = y + h^*$ is a function from $C^{(1)}$. We then denote the second integral in this equality by $I_2(h)$ and rewrite ΔI in the form

$$\Delta I = \int_a^b (F_y h + F_{y'} h')\,dx + I_2(h). \tag{3}$$

Here, as it is customarily done in the calculus of variations, the partial derivatives $\partial F/\partial y$, $\partial F/\partial y'$, $\partial^2 F/\partial y\,\partial y'$, ... are denoted by F_y, $F_{y'}$, $F_{yy'}$, ... The first integral in (3) is additive and continuous with respect to $\delta y = h$ (see functional (7) in Sec. 6.1), therefore it will be a linear functional with respect to the increment h. Let us show that the second integral in (3) is $o(\|h\|_1)$. From the definition of the norm in the space $C^{(1)}$ (see (2) in the preceding section), we have

$$\max_{x\in[a,\,b]} |h| \leqslant \|h\|_1 \quad \text{and} \quad \max_{x\in[a,\,b]} |h'| \leqslant \|h\|_1.$$

Since by the supposition, the second partial derivatives of the function $F(x, y, z)$ are continuous, there is a constant $A > 0$ such that

$$|F_{yy}| \leqslant A, \quad |F_{yy'}| \leqslant A, \quad |F_{y'y'}| \leqslant A.$$

Taking into account these and the previous inequalities, for $I_2(h)$ in (3) we get the estimate

$$|I_2(h)| = \frac{1}{2}\left|\int_a^b (F_{yy}h^2 + 2F_{yy'}hh' + F_{y'y'}h'^2)\,dx\right|$$
$$\leqslant 2A(b-a)\,\|h\|_1^2,$$

hence, the second integral is $o(\|h\|_1)$. Consequently, the *variation of functional* (2) is represented by the first integral on the right-hand side of equality (3), and it has the form

$$\delta I = \int_a^b (F_y h + F_{y'} h')\,dx. \tag{4}$$

Analogously, we can show that the variation of the functional

$$I(y_1, \ldots, y_n) = \int_a^b F(x, y_1, y_1', \ldots, y_n, y_n')\,dx \qquad (5)$$

dependent on n functions $y_1, \ldots, y_n$ and their first derivatives in the space $C^{(1)}$ has the form

$$\delta I = \int_a^b \sum_{i=1}^n (F_{y_i} h_i + F_{y_i'} h_i')\,dx. \qquad (6)$$

The variation of the functional

$$I(y) = \int_a^b F(x, y, y', \ldots, y^{(k)})\,dx \qquad (7)$$

dependent on the function and its derivatives up to the kth order inclusively in the space $C^{(k)}$ has the form

$$\delta I = \int_a^b (F_y h + F_{y'} h' + \ldots + F_{y^{(k)}} h^{(k)})\,dx. \qquad (8)$$

2. The Concept of a Weak and a Strong Extremum of a Functional. Let us first recall the definition of the extremum of a function: the point M_0 is said to be a *point of extremum of the function* $f(M)$ if there exists a certain neighbourhood of the point M_0 in which the increment of the function retains sign; if $\Delta f > 0$, then M_0 is a point of minimum, and if $\Delta f < 0$, then M_0 is a point of maximum.

Now, we are going to define the extremum of a functional.

The *functional* $I(y)$, $y \in E$, *reaches an extremum* for $y = y_0(x)$ if there is a neighbourhood of the point $y = y_0(x)$

$$P(y_0, \varepsilon) = \{y \in E, \| y - y_0 \| < \varepsilon\},$$

in which the increment $\Delta I = I(y) - I(y_0)$ does not change its sign; and if $\Delta I > 0$, then $I(y)$ reaches a *minimum* for $y = y_0$, if $\Delta I < 0$, then it reaches a *maximum*.

We shall consider functionals $I(y)$ on a set R of differentiable functions. These functions can be regarded either as elements of the space C or as elements of the space $C^{(1)}$.

Accordingly, we shall investigate functionals for an extremum in the spaces C and $C^{(1)}$.

Example 1°. Show that if the functional $I(y)$ has an extremum in the space C, then it also has an extremum in the space $C^{(1)}$.

Let y_0 extremize the functional $I(y)$ in the space C. Then, there exists an $\varepsilon > 0$ such that for all $y \in P_C(y_0, \varepsilon) = \{y \in R, \| y - y_0 \|_C < \varepsilon\}$ the increment ΔI preserves sign. Let us take the found $\varepsilon > 0$ and consider the ε-neighbourhood of the point y_0, this time with respect to the norm of the space $C^{(1)}$, that is, the neighbourhood $P_1(y_0, \varepsilon) = \{y \in R, \| y - y_0 \|_1 < \varepsilon\}$. The norm in the space $C^{(1)}$ is not less than the norm in the space C, $\| y - y_0 \|_1 \geqslant \| y - y_0 \|_C$, therefore any point $y \in P_1(y_0, \varepsilon)$ is also a point from the neighbourhood $P_C(y_0, \varepsilon)$. In other words, the ε-neighbourhood $P_C(y_0, \varepsilon)$ regarded as the set of differentiable functions is broader than the ε-neighbourhood $P_1(y_0, \varepsilon)$. For any $y \in P_C(y_0, \varepsilon)$, for $y \in P_1(y_0, \varepsilon)$ in particular, the increment of the functional ΔI preserves sign, and, consequently, y_0 extremizes the functional $I(y)$ in the space $C^{(1)}$. The converse is, generally speaking, false, that is, if y_0 extremizes the functional in the space $C^{(1)}$, then the point y_0 may be or may not be extreme in the space C. This assertion follows from the fact that if the ε-neighbourhood $P_C(y_0, \varepsilon)$ is regarded as the set of differentiable functions, then it may contain such $y \in P_C(y_0, \varepsilon)$ and $y \notin P_1(y_0, \varepsilon)$ that the increment ΔI alters its sign in the neighbourhood $P_C(y_0, \varepsilon)$.

An extremum in the space C is termed a *strong extremum*, and an extremum in the space $C^{(1)}$ a *weak extremum*. The foregoing implies that any strong extremum is also a weak extremum. The converse is, generally speaking, incorrect.

Finding a weak extremum is, as a rule, a simpler problem than finding a strong extremum. This is due to the fact that the functionals under consideration, being continuous in the space $C^{(1)}$, are rather often discontinuous in the space C. Note that finding the sufficient conditions for existence of an extremum of a functional is a very complicated problem, therefore we shall confine ourselves to considering only the necessary condition for an extremum of a functional to exist.

3. Necessary Condition for an Extremum. We shall deal with functional having a variation.

Theorem 1. *If a functional $I(y)$ reaches an extremum for $y = y_0$, then its variation vanishes for $y = y_0$, that is,*

$$\delta I|_{y=y_0} = 0. \tag{9}$$

Let the functional $I(y)$ be given in a normed space E with norm $\| y \|$ and let it attain, for definiteness, a minimum for $y = y_0$. By the definition of the minimum of a functional, its increment

$$\Delta I = I(y) - I(y_0) = I(y_0 + h) - I(y_0) > 0$$

in some ε-neighbourhood of the point y_0 is

$$P(y_0, \varepsilon) = \{y \in E, \ \| y - y_0 \| = \| h \| < \varepsilon\}.$$

By the hypothesis, there exists a variation, therefore the increment is representable in the form

$$\Delta I = \delta I(h) + o(\| h \|) > 0.$$

Let us assume that the variation $\delta I(h) \neq 0$. Since the quantity $o(\| h \|)$ is an infinitesimal of a higher order than $\| h \|$, it does not affect the sign of the increment and therefore

$$\operatorname{sgn} \Delta I(h) = \operatorname{sgn} \delta I(h).$$

By the assumption, $\Delta I(h) > 0$ in the ε-neighbourhood of the point y_0, consequently, also $\delta I(h) > 0$ in this neighbourhood.

If $y_0 + h$ belongs to the ε-neighbourhood of the point y_0, then $y_0 - h$ also belongs to this neighbourhood. The linearity of the functional implies that $\delta I(-h) = -\delta I(h)$.

Thus, the values of the functional δI for h and $(-h)$ have opposite signs, and the increment ΔI does not preserve sign in any ε-neighbourhood of the point y_0. Consequently, y_0 cannot minimize the functional. The obtained contradiction just proves the theorem.

Remark. If the functional reaches a weak extremum at the point y_0, then its variation vanishes for $y = y_0$, i.e $\delta I|_{y=y_0} = 0$ in the space $C^{(1)}$. If y_0 gives the functional a strong extremum, then it also yields a weak extremum, and therefore the equality $\delta I|_{y=y_0} = 0$ is fulfilled in both spaces: C and $C^{(1)}$. Consequently, the necessary condition for a weak

and a strong extremum is that the variation of the functional δI be equal to zero in the space $C^{(1)}$.

Note that the sufficient conditions for a weak and a strong extremum are different and, as it has already been mentioned, since their deduction involves many difficulties, we are not going to dwell on them.

4. The Fundamental Lemma of the Calculus of Variations. When solving variation problems, the following lemma turns out to be very helpful.

Lemma. *Let $\alpha(x)$ be a fixed function continuous on $[a, b]$. If for any function $h(x)$ such that $h(a) = h(b) = 0$ and its derivative continuous on $[a, b]$ the following equality is valid:*

$$\int_a^b \alpha(x) h(x)\, dx = 0, \tag{10}$$

then $\alpha(x) \equiv 0$ on (a, b).

Let us assume that $\alpha(x) \not\equiv 0$, then there is a point $\xi \in (a, b)$ such that $\alpha(\xi) \neq 0$; let for definiteness, $\alpha(\xi) > 0$. According to the continuity property, there exists an interval $(x_1, x_2) \subset (a, b)$ in which $\alpha(x) > 0$.

Consider the function

$$\tilde{h}(x) = \begin{cases} (x - x_1)^2 (x_2 - x)^2, & x \in (x_1, x_2), \\ 0, & x \notin (x_1, x_2). \end{cases} \tag{11}$$

It is obvious that $\tilde{h}(x)$ and $\tilde{h}'(x)$ are continuous on $[a, b]$ and $\tilde{h}(a) = \tilde{h}(b) = 0$. The integrand $\alpha(x)\tilde{h}(x)$ is positive for $x \in (x_1, x_2)$; by the property of integral, the inequality sign is also maintained for the integral

$$\int_a^b \alpha(x) \tilde{h}(x)\, dx = \int_{x_1}^{x_2} \alpha(x) (x - x_1)^2 (x_2 - x)^2\, dx > 0.$$

Thus, for the chosen function $\tilde{h}(x)$ relationship (10) is not fulfilled, that is, we have arrived at a contradiction with the conditions of the lemma.

Remark. The fundamental lemma of the calculus of variations also holds for a more narrow class of functions, namely, when $h(x)$ is continuous on $[a, b]$ together with the de-

rivatives up to the nth order inclusively. In this case, instead of function (11), we may consider the function

$$h(x)=\begin{cases}(x-x_1)^{2n}(x-x_2)^{2n}, & x\in(x_1,\ x_2),\\ 0, & x\notin(x_1,\ x_2).\end{cases}$$

Sec. 6.3. VARIATION PROBLEMS WITH FIXED BOUNDARIES

1. Euler's Equation. We usually distinguish between variation problems with fixed boundaries and those with moving (or free) boundaries. We shall first consider a problem with fixed boundary points since this is the simplest case. We shall regard that the integrands in the functionals under consideration are continuous and have continuous partial derivatives up to the needed order, and the functionals themselves are continuous in the space under consideration. Let us begin with the simplest variation problem for functional (2) from the preceding section, and obtain for it the necessary conditions in the form of a differential equation.

The functional

$$I(y)=\int_a^b F(x,\ y,\ y')\,dx \tag{1}$$

will be considered on the set of functions $y(x)\in C^{(1)}$ satisfying the boundary conditions

$$y(a)=A,\quad y(b)=B. \tag{2}$$

These conditions mean that the end points of permissible curves are fixed.

The following variation problem is set: *Among all functions $y(x)$ with the boundary conditions* (2), *find such conditions which extremize functional* (1). The solution of this problem will be carried out only within the framework of the necessary conditions for which purpose we shall prove the following theorem.

Theorem 1. *If the function $y=y(x)\in C^{(1)}$ satisfies conditions* (2) *and extremizes functional* (1), *then it is a solution of Euler's equation*

$$F_y-\frac{d}{dx}F_{y'}=0. \tag{3}$$

For the variation of functional (1) we take advantage of formula (4) of the preceding section

$$\delta I = \int_a^b (F_y h + F_{y'} h')\, dx.$$

Let the function $y(x)$ extremize functional (1). By the necessary condition for an extremum, the variation of a functional must equal zero, that is,

$$\delta I = \int_a^b (F_y h + F_{y'} h')\, dx = 0. \tag{4}$$

Integrating by parts, we obtain

$$\int_a^b F_{y'} h'\, dx = F_{y'} h \Big|_a^b - \int_a^b \frac{d}{dx} F_{y'} h\, dx.$$

From the boundary conditions (2) it follows that the considered increments $h(x)$ must vanish at the points $x = a$ and $x = b$, that is,

$$h(a) = h(b) = 0. \tag{5}$$

Substituting then the value of the integral $\int_a^b F_{y'} h'\, dx$ into (4), we get

$$\int_a^b \left(F_y - \frac{d}{dx} F_{y'}\right) h\, dx = 0.$$

But the function $\alpha(x) = F_y - (d/dx) F_{y'}$ is continuous on $[a, b]$ and $h(x)$ is any function continuous together with the first derivative on $[a, b]$ and satisfying conditions (5). Applying the fundamental lemma of the calculus of variations, we conclude that $\alpha(x) = F_y - (d/dx) F_{y'} \equiv 0$, i.e. the function $y(x)$ satisfies Euler's equation (3).

The functions which are solutions of Euler's equation are called *extremals*.

Theorem 1 provides only for the necessary condition for the existence of an extremum of a functional. But frequently,

the existence of an extremum is clear from physical considerations. In this case Euler's equation solves a variation problem completely.

Euler's equation (3) plays a fundamental role in the calculus of variations, as a whole. It represents a second-order differential equation. Its solution depends on two arbitrary constants. In Cauchy's problem, these arbitrary constants were found from the initial conditions. Here we have another problem for differential equations—a boundary-value problem in which the arbitrary constants are found from the boundary conditions. Thus, for Euler's equation (3) arbitrary constants are found from conditions (2). In the general case, equation (3) is not solvable by quadratures.

2. Particular Cases of Euler's Equation. Let us consider several cases when equation (3) allows the reduction of order and is reduced to a first-order differential equation.

(1) *The function $F(x, y, y')$ does not contain y explicitly, that is, $F(x, y, y') = F(x, y')$.* In this case $F_y = 0$, and Euler's equation takes the form

$$\frac{\mathrm{d}}{\mathrm{d}x} F_{y'} = 0.$$

Hence we find that

$$F_{y'} = c = \text{const}, \tag{6}$$

that is, we have a first-order differential equation not containing y explicitly.

Example 1°. A material point moves from the point $A(1, 0)$ to the point $B(2, 1)$ with velocity $v = x$. Find the curve of the shortest motion time.

The time spent by the material point to cover the arc length of the curve $y = y(x)$ is determined with the aid of the integral

$$t(y) = \int_1^2 \frac{\sqrt{1+y'^2}}{x}\,\mathrm{d}x$$

representing a functional of form (1) in which the considered curves $y(x)$ satisfy the conditions $y(1) = 0$ and $y(2) = 1$. The integrand $F(x, y, y') = \dfrac{\sqrt{1+y'^2}}{x}$ is independent of y,

therefore from (6) we have the equalities

$$F_{y'} = \frac{y'}{x\sqrt{1+y'^2}} = \frac{1}{c}.$$

Determining now y'

$$y' = \pm \frac{x}{\sqrt{c^2 - x^2}}$$

and integrating, we find the extremals

$$y = c_1 \pm \sqrt{c^2 - x^2}, \text{ or } (y - c_1) + x^2 = c^2.$$

From the boundary conditions $y(1) = 0$ and $y(2) = 1$ for determining c and c_1 we obtain the system

$$c_1^2 - 1 = c^2, \quad (1 - c_1)^2 + 4 = c^2.$$

Hence we find that $c_1 = 2$, $c^2 = 5$, and the equation of the required extremal is a circle $x^2 + (y - 2)^2 = 5$ of radius $\sqrt{5}$ centred at the point (0, 2).

From physical considerations it is clear that there is no maximum for the time of motion along distinct curves, and the function $y = 2 - \sqrt{5 - x^2}$ provides a minimum for the functional.

(2) *The function $F(x, y, y')$ does not contain x explicitly, that is, $F(x, y, y') = F(y, y')$.* Let us multiply both sides of equation (3) by y':

$$y'F_y - y'\frac{\mathrm{d}}{\mathrm{d}x}F_{y'} = 0,$$

and write this equality in the form

$$y'F_y + y''F_{y'} - \left(y''F_{y'} + y'\frac{\mathrm{d}}{\mathrm{d}x}F_{y'}\right) = 0.$$

The function $F = F(y, y')$ does not depend explicitly on x, therefore its total derivative with respect to x is representable by the expression

$$\frac{\mathrm{d}F}{\mathrm{d}x} = y'F_y + y''F_{y'}.$$

Besides, we have the equality

$$y''F_{y'} + y'\frac{\mathrm{d}}{\mathrm{d}x}F_{y'} = \frac{\mathrm{d}}{\mathrm{d}x}(y'F_{y'}),$$

therefore Euler's equation takes the form

$$\frac{d}{dx}(F - y'F_{y'}) = 0.$$

Hence we find that

$$F - y'F_{y'} = C. \tag{7}$$

Thus, we have obtained a first-order equation not containing x explicitly.

Example 2°. Among the curves connecting two points (x_1, y_1) and (x_2, y_2) find the one whose rotation about the axis of abscissas generates a surface of minimum area (the minimum-surface-of-revolution problem).

The area of a surface of revolution about the x-axis is given by the functional

$$I(y) = 2\pi \int_{x_1}^{x_2} y \sqrt{1 + y'^2}\, dx$$

the permissible curves $y(x)$ satisfying the conditions $y(x_1) = y_1$ and $y(x_2) = y_2$. The integrand is independent of x, therefore for Euler's equation we may take advantage of formula (7) which in this case takes the form

$$y\sqrt{1+y'^2} - \frac{yy'^2}{\sqrt{1+y'^2}} = c, \quad \text{or} \quad y = c\sqrt{1+y'^2}.$$

After elementary transformations we obtain the equation

$$\frac{c\,dy}{\sqrt{y^2 - c^2}} = dx.$$

Integrating this equation, we get

$$\ln\left|\frac{y}{c} + \sqrt{\left(\frac{y}{c}\right)^2 - 1}\right| = x + c_1.$$

Solving the obtained equality with respect to y, we arrive at the equation of a catenary:

$$y = c\cosh(x + c_1).$$

The constants c and c_1 are found from the system

$$y_1 = c\cosh(x_1 + c_1), \quad y_2 = c\cosh(x_2 + c_1),$$

which may have one, two, or no solutions.

The function $F(x, y, y')$ is independent of y', that is, $F = F(x, y)$. In this case Euler's equation takes the form

$$F_y = 0, \tag{8}$$

which is not a differential equation but a final equation defining one or several curves. If among these curves there are curves satisfying condition (8), then they will be extremals.

Example 3°. Find the extremals of the functional

$$I(y) = \int_0^1 (x-y)^2\,dy \quad \text{if} \quad y(0)=1, \quad y(1)=2.$$

For the case under consideration, equation (8) is written in the form

$$F_y = -2(x-y) = 0.$$

Its solution $y = x$ does not satisfy the boundary conditions, and, consequently, is not an extremal.

The following theorem presents a particular case of the variation problem ((1), (2)) frequently encountered in applications when the extremal realizes the minimum of functional (1).

Theorem 2. *Let the functions $p(x)$, $p'(x)$, $q(x)$, and $f(x)$ be continuous on the interval $[a, b]$ and, in addition, $p(x) > 0$, $q(x) \geqslant 0$. If $y(x)$ is an extremal of the functional*

$$I(y) = \int_a^b [p(x)\,y'^2 + q(x)\,y^2 + 2yf(x)]\,dx \tag{9}$$

and satisfies the boundary conditions (2): $y(a) = A$, $y(b) = B$, *then it realizes the absolute minimum of functional* (9), *i.e. for any other permissible curve $\tilde{y} = \tilde{y}(x)$ the inequality $I(\tilde{y}) > I(y)$ is fulfilled.*

Let $y = y(x)$ be an extremal of functional (9); then it satisfies Euler's equation $(d/dx)(py') - qy - f = 0$ and the boundary conditions $y(a) = A$, $y(b) = B$. Further, let $\tilde{y}(x)$ be a permissible curve in the variation problem ((9), (2)); in particular, it satisfies the boundary conditions $\tilde{y}(a) = A$, $\tilde{y}(b) = B$. Let us introduce the notation $h(x) = \tilde{y}(x) - y(x)$, then $h(a) = h(b) = 0$. We find now the

increment of functional (9):

$$\Delta I = I(y+h) - I(y)$$

$$= \int_a^b [p(y'+h')^2 + q(y+h)^2 + 2(y+h)f]\,dx$$

$$- \int_a^b (py'^2 + qy^2 + 2fy)\,dx = 2\int_a^b (py'h' + qyh + fh)\,dx$$

$$+ \int_a^b (ph'^2 + qh^2)\,dx.$$

Integrating by parts, we transform the integral:

$$\int_a^b py'h'\,dx = py'h\Big|_a^b - \int_a^b h\frac{d}{dx}(py')\,dx = -\int_a^b h\frac{d}{dx}(py')\,dx$$

and substitute it into ΔI, then

$$\Delta I = 2\int_a^b \left[qy - \frac{d}{dx}(py') + f\right] h\,dx + \int_a^b (ph'^2 + qh^2)\,dx.$$

But y is a solution of the equation $(d/dx)(py') - qy - f = 0$, therefore

$$\Delta I = \int_a^b (ph'^2 + qh^2)\,dx.$$

Since $p > 0$, $q \geqslant 0$, $\Delta I > 0$. From the definition of the minimum of a functional it follows that $y(x)$ realizes the absolute minimum of functional (9).

Theorem 2 can be used as a sufficient condition when investigating functional (9) for an extremum—the solution of Euler's equation satisfying conditions (2) minimizes functional (9).

Example 4°. Investigate for an extremum the functional

$$I(y) = \int_1^2 (x^2y'^2 + 12y^2)\,dx, \quad y(1) = 1, \quad y(2) = 8.$$

Euler's equation for the given functional has the form

$$x^2y'' + 2xy' - 12y = 0.$$

In the theory of differential equations, linear equations of such type are called Euler's equations. We find its solution in the form $y = x^\lambda$. On finding the derivatives $y' = \lambda x^{\lambda-1}$ and $y'' = \lambda(\lambda - 1)x^{\lambda-2}$, we substitute them into Euler's equation to obtain

$$x^\lambda(\lambda^2 + \lambda - 12) = 0.$$

For determining λ, we have the characteristic equation $\lambda^2 + \lambda - 12 = 0$ whose roots are: $\lambda_1 = 3$ and $\lambda_2 = -4$. The general solution of Euler's equation has the form

$$y = c_1x^3 + c_2x^{-4}.$$

To determine the constants c_1 and c_2, from the boundary conditions $y(1) = 1$ and $y(2) = 8$ we obtain the system

$$c_1 + c_2 = 1, \quad 8c_1 + \frac{c_2}{16} = 8.$$

Hence we find: $c_1 = 1$ and $c_2 = 0$. Consequently, $y = x^3$ is an extremal of the given functional. For the example under consideration, Theorem 2 ($p(x) = x^2 > 0$, $q(x) = 12 > 0$ on $[1, 2]$) is applicable, therefore the extremal $y = x^3$ realizes the minimum of the functional.

3. Variation Problem for a Functional Depending on n Functions. The problem is formulated in the following way: find the necessary conditions for an extremum of the functional

$$I(y_1, \ldots, y_n) = \int_a^b F(x, y_1, y_1', \ldots, y_n, y_n')\,dx \qquad (10)$$

dependent on n functions $y_1, y_2, \ldots, y_n \in C^{(1)}$ satisfying the boundary conditions

$$y_i(a) = A_i, \quad y_i(b) = B_i, \quad i = 1, 2, \ldots, n. \qquad (11)$$

Within the bounds of the necessary condition, the answer is given by the following theorem.

Theorem 3. *If a system of linearly independent functions $y_1(x), y_2(x), \ldots, y_n(x)$ satisfying conditions* (11) *extremizes*

functional (10), *then it is a solution of the system of Euler's differential equations*

$$F_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x} F_{y_i'} = 0, \quad i = 1, 2, \ldots, n. \tag{12}$$

As it was shown in formula (6) of the preceding section, the variation of functional (10) is written in the form

$$\delta I = \int_a^b \sum_{i=1}^{n} (F_{y_i} h_i + F_{y_i'} h_i') \, \mathrm{d}x.$$

All the increments $h_i(x)$ are independent, therefore one of them, say h_ν, is chosen arbitrarily with the boundary conditions observed, all the rest being regarded as equal to zero, that is,

$$h_1(x) \equiv \ldots \equiv h_{\nu-1}(x) \equiv h_{\nu+1}(x) \equiv \ldots \equiv h_n(x) \equiv 0.$$

From the necessary condition for an extremum of a functional we may write that

$$\delta I = \int_a^b (F_{y_\nu} h_\nu + F_{y_\nu'} h_\nu') \, \mathrm{d}x = 0. \tag{13}$$

Hence, we have a simplest variation problem and may apply to it Theorem 1, according to which the function $y_\nu(x)$ must satisfy the equation $F_{y_\nu} - (\mathrm{d}/\mathrm{d}x) F_{y_\nu'} = 0$. But equality (12) can be written for any $\nu = 1, 2, \ldots, n$, consequently, each of the functions $y_1(x)$, $y_2(x)$, $\ldots$ $\ldots$, $y_n(x)$ simultaneously satisfies Euler's equation (3), that is, their collection is the solution of the system of Euler's equations (12). The system consists of n second-order equations. Its general solution contains $2n$ arbitrary constants which are found from the boundary conditions (10).

Example 5°. Find the extremals of the functional

$$I(y, z) = \int_2^3 (xy'^2 + z'^2 + xy'z') \, \mathrm{d}x$$

satisfying the boundary conditions $y(2) = \ln 3$, $y(3) = \ln 3$, $z(2) = \ln 2$, and $z(3) = 0$.

The integrand does not contain y and z explicitly, therefore the system of Euler's equations has the form

$$2xy' + xz' = c_1, \quad 2z' + xy' = c_2.$$

From this system we find the derivatives y' and z':

$$y' = \frac{c_2 x - 2c_1}{x(x-4)}, \quad z' = \frac{c_1 - 2c_2}{x-4}.$$

Integrating this system to within arbitrary constants, we find the functions

$$y = \frac{2c_2 + c_1}{2} \ln|x-4| - \frac{c_1}{2} \ln|x| + c_3,$$
$$z = (c_1 - 2c_2) \ln|x-4| + c_4.$$

The arbitrary constants are determined from the system

$$2c_2 \ln 2 + c_3 = \ln 3, \quad -c_1 \ln 3 + 2c_3 = \ln 3,$$
$$(c_1 - 2c_2) \ln 2 + c_4 = \ln 2, \quad c_4 = 0:$$
$$c_1 = 1, \quad c_2 = 0, \quad c_3 = \ln 3, \quad c_4 = 0.$$

Hence, the desired extremal has the form

$$y = \frac{1}{2} \ln \frac{9|x-4|}{|x|}, \quad z = \ln|x-4|.$$

4. Functionals Dependent on Higher-order Derivatives. Let us consider a variation problem for the functional

$$I(y) = \int_a^b F(x, y, y', y'', \ldots, y^{(k)})\, dx \tag{14}$$

given on the set of functions belonging to the space $C^{(k)}$ and satisfying the boundary conditions

$$y(a) = A, \ y'(a) = A_1, \ y''(a) = A_2, \ldots, y^{(k-1)}(a) = A_{k-1},$$
$$y(b) = B, \ y'(b) = B_1, \ y''(b) = B_2, \ldots, y^{(k-1)}(b) = B_{k-1}. \tag{15}$$

We shall hold that the function $F(x, y, z_1, \ldots, z_k)$ has continuous partial derivatives up to the order $k+1$ inclu-

sively, and obtain a necessary condition for the existence of an extremum of functional (14) in the case of the fixed boundaries (15).

Theorem 3. *If the function* $y(x) \in C^{(k)}$ *satisfying the boundary conditions* (15) *extremizes functional* (14), *then it is a solution of the Euler-Poisson differential equation*

$$F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'} + \frac{\mathrm{d}^2}{\mathrm{d}x^2} F_{y''} - \ldots + (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} F_{y^{(k)}} = 0. \quad (16)$$

According to formula (8) derived in Sec. 6.2, the variation of functional (14) is written in the form

$$\delta I = \int_a^b (F_y h + F_{y'} h' + \ldots + F_{y^{(k)}} h^{(k)})\, \mathrm{d}x.$$

Integrating by parts, we have

$$\delta I = \int_a^b F_y h\, \mathrm{d}x + F_{y'} h \Big|_a^b - \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} F_{y'} h\, \mathrm{d}x + F_{y''} h' \Big|_a^b$$

$$- \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} F_{y''} h'\, \mathrm{d}x + \ldots + F_{y^{(k)}} h^{(k-1)} \Big|_a^b - \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} F_{y^{(k)}} h^{(k-1)}\, \mathrm{d}x.$$

From the boundary conditions (15) it follows that the increments $h(x)$ satisfy the conditions

$$h(a) = h(b) = h'(a) = h'(b) = \ldots = h^{(k-1)}(a) = h^{(k-1)}(b) = 0, \quad (17)$$

therefore

$$\delta I = \int_a^b \left(F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'}\right) h\, \mathrm{d}x - \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} F_{y''} h'\, \mathrm{d}x - \ldots - \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} F_{y^{(k)}} h^{(k-1)}\, \mathrm{d}x.$$

Integrating by parts the second and consequent integrals a necessary number of times and using conditions (17), we get

$$\delta I = \int_a^b \left(F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'} + \frac{\mathrm{d}^2}{\mathrm{d}x^2} F_{y''} - \ldots + (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} F_{y^{(k)}}\right) h\, \mathrm{d}x.$$

Applying a necessary condition for an extremum of a functional, we conclude that $\delta I = 0$. From the remark to the fundamental lemma of the calculus of variations it follows that a function extremizing functional (14) must satisfy the Euler-Poisson equation (16) having the order $2k$. The general solution depends on $2k$ arbitrary constants which can be determined from the boundary conditions (15).

Example 6°. Find the extremals of the functional

$$I(y) = \int_0^{\pi/2} (y''^2 - y^2 + x^2)\, dx$$

for the conditions: $y(0) = 1$, $y'(0) = 0$, $y(\pi/2) = 0$, and $y'(\pi/2) = -1$.

Let us write the Euler-Poisson equation for the given functional

$$-2y + 2y^{IV} = 0.$$

The general solution of this equation is written in the form

$$y = c_1 e^x + c_2 e^{-x} + c_3 \cos x + c_4 \sin x.$$

Using the boundary conditions for the constants c_1, c_2, c_3, c_4, we obtain the system of equations

$$c_1 + c_2 + c_3 = 1, \quad c_1 - c_2 + c_4 = 0,$$
$$c_1 e^{\pi/2} + c_2 e^{-\pi/2} + c_4 = 0, \quad c_1 e^{\pi/2} - c_1 e^{-\pi/2} - c_3 = -1.$$

Solving this system, we find: $c_1 = c_2 = c_4 = 0$ and $c_3 = 1$. Thus, the extremal of the given functional is the function $y = \cos x$.

Analogously, we may prove that the extremals of the functional

$$I(y_1, \ldots, y_n) = \int_a^b F(x, y_1, y_1', \ldots, y_1^{(k)}, \ldots, y_n, y_n', \ldots, y_n^{(k)})\, dx$$

dependent on several functions and their derivatives satisfying the boundary conditions

$$y_i(a) = A_i,\ y_i(b) = B_i,\ y_i'(a) = A_{i_1},\ y_i'(b) = B_{i_2}, \ldots,$$

$$y_i^{(k-1)}(a) = A_{i,\,k-1},\ y_i^{(k-1)}(b) = B_{i,\,k-1},\ i = 1, 2, \ldots, n,$$

are a solution of the system of the Euler-Poisson equations

$$F_{y_i} - \frac{d}{dx} F_{y_i'} + \ldots + (-1)^k \frac{d^k}{dx^k} F_{y_i^{(k)}} = 0, \quad i = 1, 2, \ldots, n.$$

Sec. 6.4. VARIATION PROBLEMS INVOLVING A CONDITIONAL EXTREMUM

1. Variation Problems with Final Constraints (or Accessory Conditions). Variation problems involving a conditional extremum embrace problems in which it is required to find the extremum of a functional if in addition to boundary conditions, the conditions of another type, so-called *constraint conditions*, are imposed on permissible curves. Constraint conditions can be specified by systems of finite or differential equations.

Consider the variation problem for the functional

$$I(y_1, \ldots, y_n) = \int_a^b F(x, y_1, y_1', \ldots, y_n, y_n')\, dx \qquad (1)$$

defined on a set of functions belonging to the space $C^{(1)}$ with the boundary conditions

$$y_i(a) = A_i,\ y_i(b) = B_i,\ i = 1, 2, \ldots, n. \qquad (2)$$

Besides, let us assume that the functions $y_1, y_2, \ldots, y_n$ satisfy k constraint equations

$$\varphi_j(x, y_1, y_2, \ldots, y_n) = 0, \quad j = 1, 2, \ldots, k,\ k < n. \qquad (3)$$

All functions φ_j are regarded to be independent and continuous together with the partial derivatives with respect to all variables.

The variation problem ((1), (2), (3)) is known as a *variation problem involving a conditional extremum with final con-*

straints. Let us consider how this problem is solved by the method of undetermined Lagrangian multipliers.

Theorem 1. *If the system of functions*

$$y_1(x),\ y_2(x),\ \ldots,\ y_n(x) \tag{4}$$

satisfying the boundary conditions (2) *and constraint conditions* (3) *maximizes or minimizes functional* (1), *then there exist functions* $\lambda_j(x), j = 1, 2, \ldots, k$, *such that the system of functions* (4) *is the extremal of the functional*

$$I^*(y_1, \ldots, y_n) = \int_a^b \left[F + \sum_{j=1}^{k} \lambda_j(x)\,\varphi_j(x)\right] dx. \tag{5}$$

The unknown functions $\lambda_j(x)$ are called *Lagrangian multipliers.*

From formula (6) given in Sec. 6.2, by integrating by parts, we obtain the following equality for the variation of functional (1):

$$\delta I = \int_a^b \left(\sum_{i=1}^{n} F_{y_i} - \frac{d}{dx} F_{y_i'}\right) h_i\, dx. \tag{6}$$

According to the necessary condition for an extremum of a functional for the family of functions (4) we may write that $\delta I = 0$. The functions $y_1, \ldots, y_n$ are subject to (3), therefore the increments h_i, $i = 1, 2, \ldots, n$, are no longer arbitrary. This means that it is impossible to put all h_i, but one, equal to zero, and, consequently, it is impossible to apply the fundamental lemma of the calculus of variations. Let us transform variation (6). From the constraint equations we have:

$$\varphi_j(x,\ y_1 + h_1,\ \ldots,\ y_n + h_n) = 0,$$
$$\varphi_j(x,\ y_1,\ \ldots,\ y_n) = 0,\ j = 1,\ 2,\ \ldots,\ k.$$

Applying Taylor's formula to the difference of these equations, we obtain

$$0 = \varphi_j(x,\ y_1 + h_1,\ \ldots,\ y_n + h_n) - \varphi_j(x,\ y_1,\ \ldots,\ y_n)$$
$$\sum_{i=1}^{n} \frac{\partial \varphi_j}{\partial y_i} h_i + o\left(\sum_{i=1}^{n} \|h_i\|_1\right).$$

Hence, it follows (to within an infinitesimal) that

$$\sum_{i=1}^{n} \frac{\partial \varphi_j}{\partial y_i} h_i = 0, \quad j = 1, 2, \ldots, k.$$

Multiplying these equalities by certain functions $\lambda_j(x)$ and summing with respect to all j, we get the following identity:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} h_i = 0;$$

integrating it between the limits from a to b, we conclude that the below integral is equal to zero:

$$\int_a^b \sum_{j=1}^{k} \sum_{i=1}^{n} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} h_i \, dx = 0.$$

Adding this integral to variation (6) and taking into consideration that $\delta I = 0$, we write the necessary condition for an extremum of a functional in the form

$$\delta I = \int_a^b \sum_{i=1}^{n} \left(F_{y_i} - \frac{d}{dx} F_{y_i'} + \sum_{j=1}^{k} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} \right) h_i \, dx = 0. \quad (7)$$

By the hypothesis, the functions $\varphi_j(x, y_1, \ldots, y_n)$, $j = 1, \ldots, k$, are independent, therefore there exists a Jacobian of the order k which is different from zero. Let, for instance,

$$\frac{\partial(\varphi_1, \ldots, \varphi_k)}{\partial(y_1, \ldots, y_k)} = \begin{vmatrix} \frac{\partial \varphi_1}{\partial y_1} & \cdots & \frac{\partial \varphi_1}{\partial y_k} \\ \cdots & \cdots & \cdots \\ \frac{\partial \varphi_k}{\partial y_1} & \cdots & \frac{\partial \varphi_k}{\partial y_k} \end{vmatrix} \neq 0.$$

Hence it follows that, firstly, the functions $y_1, \ldots, y_k$ are expressed in terms of the functions $y_{k+1}, y_{k+2}, \ldots, y_n$, which are independent, and, secondly, the determinant of the system of equations

$$\sum_{j=1}^{k} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} = \frac{d}{dx} F_{y_i'} - F_{y_i}, \quad i = 1, 2, \ldots, k, \quad (8)$$

which is the chosen Jacobian, is different from zero. Consequently, system (8) has the solution $\lambda_1(x)$, $\lambda_2(x)$, ..., $\lambda_k(x)$. We substitute the found functions $\lambda_j(x)$ into equation (7) and reduce it to the form

$$\int_a^b \sum_{i=k+1}^{n} \left\{ F_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x} F_{y_i'} + \sum_{j=1}^{k} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} \right\} h_i \, \mathrm{d}x = 0.$$

In this equality, the increments h_{k+1}, h_{k+2}, ..., h_n turn out to be independent. Therefore, setting all h_i, but one, equal to zero and applying the fundamental lemma of the calculus of variations, we find that the function y_i satisfy the equations

$$F_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x} F_{y_i'} + \sum_{j=1}^{k} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} = 0, \qquad i = k+1, \ldots, n.$$

Uniting these equations with equations (8), we obtain that the functions $y_i(x)$ and $\lambda_j(x)$ satisfy the system of equations

$$F_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x} F_{y_i'} + \sum_{j=1}^{k} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} = 0 \tag{9}$$

for all $i = 1, 2, \ldots, n$. The system of differential equations (9) is the system of Euler's equations for functional (5). Consequently, the curves $y_1, y_2, \ldots, y_n$ are the extremals of functional (5).

From this theorem it is possible to draw the following conclusion. In order to find the extremals of the variation problem involving a conditional extremum ((1)-(3)), one has to find the extremals of functional (5) satisfying the boundary conditions (2) and constraint equations (3). Thus, in the variation problem involving a conditional extremum, it is desired to find n extremals $y_1(x)$, $y_2(x)$, ..., $y_n(x)$ and k Lagrangian multipliers $\lambda_1(x)$, $\lambda_2(x)$, ..., $\lambda_k(x)$. The problem is thus reduced to solving system (9) of n second-order differential equations

$$F_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x} F_{y_i'} + \sum_{j=1}^{k} \lambda_j(x) \frac{\partial \varphi_j}{\partial y_i} = 0, \quad i = 1, 2, \ldots, n,$$

and k constraint equations (3)

$$\varphi_j(x, y_1, \ldots, y_n) = 0;$$

$2n$ arbitrary constants are found from the boundary conditions (2).

Remark. Theorem 1 solves the problem involving a conditional extremum only in the limits of the necessary conditions, that is, all extremals for an unconditional extremum of functional (5) will be the extremals for a conditional extremum of functional (1) satisfying the boundary conditions (2) and constraint equations (3). But the theorem gives no answer to the question whether all solutions concerning the conditional extremum of the problem ((1)-(3)) can be found using this method. Since this question is rather complicated, it is not dealt with in this study aid.

Example 1°. Find the geodesic on the surface $\varphi(x, y, z) = 0$ joining the points $A(x_0, y_0, z_0)$ and $B(x_1, y_1, z_1)$*.

We shall seek the equation of the desired curve in the form of the intersection of the cylindrical surfaces $y = y(x)$ and $z = z(x)$. In this case the length of the curve is computed by the formula

$$I(y, z) = \int_{x_0}^{x_1} \sqrt{1 + y'^2 + z'^2}\, dx.$$

The curve lies on the surface, therefore the relationship $\varphi(x, y(x), z(x)) = 0$ holds true. In order to solve the formulated problem, let us form an auxiliary functional:

$$I^*(y, z, \lambda) = \int_{x_0}^{x_1} \left[\sqrt{1 + y'^2 + z'^2} + \lambda(x)\, \varphi(x, y, z)\right] dx.$$

The system of Euler's equations for this functional has the form

$$\begin{aligned} \lambda(x)\, \varphi_y - \frac{d}{dx} \frac{y'}{\sqrt{1 + y'^2 + z'^2}} &= 0, \\ \lambda(x)\, \varphi_z - \frac{d}{dx} \frac{z'}{\sqrt{1 + y'^2 + z'^2}} &= 0. \end{aligned} \tag{10}$$

* The shortest lines joining any two points on a surface are the *geodesics of that surface*. In a plane, the geodesics are Euclidean straight lines. On a sphere, the geodesics are great circles.

From this system and the constraint equation $\varphi(x, y, z) = 0$ we determine the desired functions $y(x)$, $z(x)$ and auxiliary multiplier $\lambda(x)$. The two arbitrary constants are found from the conditions that the extremals pass through the points A and B.

Let us find, in particular, the geodesic of the cylinder $x^2 + y^2 = 4$, joining the points A (0, 2, 0) and B $(1, \sqrt{3}, \pi)$. We write an auxiliary functional:

$$I^*(y, z, \lambda) = \int_0^1 [\sqrt{1 + y'^2 + z'^2} + \lambda(x)(x^2 + y^2 - 4)]\, dx.$$

Using system (10) and the constraint equation, we see that the desired geodesic satisfies the system of equations

$$\frac{z'}{\sqrt{1 + y'^2 + z'^2}} = \frac{1}{c},$$

$$2\lambda(x)\, y - \frac{d}{dx}\frac{y'}{\sqrt{1 + y'^2 + z'^2}} = 0,$$

$$x^2 + y^2 = 4.$$

From the first equation we find $z' = c_1\sqrt{1 + y'^2}$, where $c_1 = (c^2 - 1)^{-1/2}$. From the third equation we have $y' = \pm x/\sqrt{4 - x^2}$, therefore $\sqrt{1 + y'^2} = 2/\sqrt{4 - x^2}$. Substituting the value $\sqrt{1 + y'^2}$ into the expression for z', we obtain

$$z' = \frac{c_1}{\sqrt{4 - x^2}} \quad \text{and} \quad z = c_1 \arcsin\left(\frac{x}{2}\right) + c_2.$$

From the boundary conditions $z(0) = 0$ and $z(1) = \pi$ we have $c_1 = 6$ and $c_2 = 0$. We take into consideration that $y = y(x)$ is determined from the third equation, then we may not compute the auxiliary function $\lambda(x)$ from the second equation.

Thus, the equation of the geodesic on the cylinder will be written in the form

$$x^2 + y^2 = 4, \quad z = 6 \arcsin\frac{x}{2},$$

or in perimetric form

$$x = 2 \sin t,\ y = 2 \cos t,\ z = 6t,$$

this is the equation of a helix.

Remark. If we succeed in expressing some variables from the constraint equations (3) in terms of independent variables, then the problem of finding a conditional extremum of a functional is reduced to the problem of determining an unconditional extremum. Thus, in Example 1, by substituting the value of y'^2 from the equation $x^2 + y^2 = 4$ into the functional $l = \int_{x_0}^{x_1} \sqrt{1 + y'^2 + z'^2}\, dx$, we get a problem concerning an unconditional extremum for the function $z = z(x)$.

2. The Problem with Differential Constraints. The statement of Theorem 1 remains true in broader suppositions, when constraint equations are stipulated by differential equations

$$\psi_j(x,\ y_1,\ y_1',\ \ldots,\ y_n,\ y_n') = 0,$$
$$j = 1, 2, \ldots, m,\ m \leqslant n. \quad (11)$$

We shall assume that there exists a nonzero Jacobian of the order m, for instance,

$$\frac{\partial(\psi_1, \psi_2, \psi_m)}{\partial(y_1', y_2', \ldots, y_m')} = \begin{vmatrix} \frac{\partial \psi_1}{\partial y_1'} & \cdots & \frac{\partial \psi_1}{\partial y_m'} \\ \cdots & \cdots & \cdots \\ \frac{\partial \psi_m}{\partial y_1'} & \cdots & \frac{\partial \psi_m}{\partial y_m'} \end{vmatrix} \neq 0.$$

Then, from the system of differential equations (11), it is possible to determine the functions $y_1, \ldots, y_m$ in terms of the independent functions $y_{m+1}, y_{m+2}, \ldots, y_n$.

Given below without proof is a theorem pertaining to this case which is the analogue of Theorem 1.

Theorem 2. *If the system of functions* $y_1(x), \ldots, y_n(x)$ *satisfying the constraint equations* (11) *and boundary conditions* (2) *minimizes or maximizes functional* (1), *then there exist functions* $\lambda_1(x), \lambda_2(x), \ldots, \lambda_m(x)$ *such that the functions* $y_1(x), \ldots, y_n(x)$ *are the extremals of the func-*

tional

$$I^*(y_1, \ldots, y_n) = \int_a^b \left(F + \sum_{j=1}^{m} \lambda_j(x)\,\varphi_j\right) dx. \qquad (12)$$

3. The Isoperimetric Problem. This problem received its name in connection with the aid of finding a closed curve of a given length (perimeter) bounding a maximum area.

If the curve is represented parametrically

$$x = x(t), \quad y = y(t), \quad \alpha \leqslant t \leqslant \beta,$$

then the problem is reduced to finding an extremum of the functional expressing the area bounded by a closed curve

$$I(x, y) = \int_\alpha^\beta xy'\,dt$$

provided that the length of this curve is constant:

$$l(x, y) = \int_\alpha^\beta \sqrt{x'^2 + y'^2}\,dt = l = \text{const}.$$

The length of the curve is also a functional. Thus, in the problem under consideration* it is required to find the extremum of one functional provided that the other functional retains a constant value.

In the general case, in the isoperimetric problem it is desired to find the extremum of the functional

$$I(y_1, \ldots, y_n) \cdot \int_a^b F(x, y_1, y_1', \ldots, y_n, y_n')\,dx, \qquad (1)$$

when the constraint conditions are given in the form of functionals attaining the defined values

$$l_j = \int_a^b F_j(x, y_1, y_1', \ldots, y_n, y_n')\,dx,$$

$$j = 1, 2, \ldots, m, \quad m \leqslant n. \qquad (13)$$

* Problems of such type were solved even in ancient Greece, but their variation character was justified only at the end of the seventeenth century by L. Euler.

The Jacobian $\partial (F_1, \ldots, F_m)/\partial (y_1, \ldots, y_m)$ is assumed to be not equal to zero. We shall consider the isoperimetric problem for the case of fixed boundaries for which purpose it is required that the boundary conditions (2) be fulfilled:

$$y_i (a) = A_i, \quad y_i (b) = B_i, \quad i = 1, 2, \ldots, n. \tag{2}$$

Theorem 3. *If the system of curves* $y_1 (x), \ldots, y_n (x)$ *satisfying the constraint conditions* (13) *and boundary conditions* (2) *minimizes or maximizes functional* (1), *then there exist constants* $\lambda_1, \ldots, \lambda_m$ *such that the functions* $y_1 (x), \ldots, y_n (x)$ *are the extremals of the functional*

$$I^* (y_1, \ldots, y_n) = \int_a^b \left(F + \sum_{j=1}^{m} \lambda_j F_j \right) dx. \tag{14}$$

Let us reduce the isoperimetric problem to the problem of finding a conditional extremum, given the differential constraint equations (11). To this end, we introduce the auxiliary functions

$$z_j (x) = \int_a^x F_j (x, y_1, y_1', \ldots, y_n, y_n') \, dx, \quad j = 1, \ldots, m,$$

which satisfy the boundary conditions

$$z_j (a) = 0, \; z_j (b) = l_j, \; j = 1, \ldots, m.$$

Instead of constraint conditions (13), we are going to consider the constraints specified by the differential equations $z_j' (x) - F_j = 0$, $j = 1, \ldots, m$. By virtue of Theorem 2, there are functions $\lambda_1 (x), \ldots, \lambda_m (x)$ such that $y_1 (x), \ldots, y_n (x)$ are the extremals of functional (14) having in this case the form

$$I^{**} = \int_a^b \left(F - \sum_{j=1}^{n} \lambda_j (x) (z_j' - F_j) \right) dx$$

$$= \int_a^b \Phi (x, y_1, y_1', \ldots, y_n, y_n', z_1', z_m') \, dx.$$

Let us write the system of Euler's equations for this functional:

$$\begin{aligned} &\Phi_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x}\Phi_{y_i'} = 0, \quad i = 1, \ldots, n, \\ &\Phi_{z_j} - \frac{\mathrm{d}}{\mathrm{d}x}\Phi_{z_j'} = 0, \quad j = 1, \ldots, m. \end{aligned} \tag{15}$$

We have $\Phi = F - \sum_{j=1}^{n} \lambda_j(x)(z_j' - F_j)$, therefore the partial derivatives $\Phi_{z_j} = 0$, $\Phi_{z_j'} = -\lambda_j(x)$ and from m last equations it follows that

$$\frac{\mathrm{d}}{\mathrm{d}x}\lambda_j(x) = 0 \quad \text{or} \quad \lambda_j(x) = \lambda_j = \text{const}, \quad j = 1, \ldots, m,$$

and system (15) takes the form

$$F_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x}F_{y_i'} + \sum_{j=1}^{m} \lambda_j \left((F_j)_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x}(F_j)_{y_i'} \right) = 0,$$

1248 $i = 1, \ldots, n.$ (16)

The last system is a system of Euler's equations for functional (14), and therefore $y_1(x), \ldots, y_n(x)$ are the extremals of this functional.

The solution of the isoperimetric problem is reduced to solving system (16) consisting of n Euler's differential equations. The solution of this system contains $2n$ arbitrary constants $c_1, \ldots, c_{2n}$ and m Lagrangian multipliers $\lambda_1, \lambda_2, \ldots, \lambda_m$. They are found from $2n$ boundary conditions (2) and m constraint conditions (13).

Example 2°. Find the curve passing through the points $A(-1, 0)$ and $B(1, 0)$, having length π and such that the area enclosed between the curve and a segment of the x-axis is maximum.

Note that the area and arc length are expressed by the integrals

$$S(y) = \int_{-1}^{1} y(x)\,\mathrm{d}x \quad \text{and} \quad l(y) = \int_{-1}^{1} \sqrt{1 + y'^2}\,\mathrm{d}x,$$

respectively, and the desired function $y(x)$ must satisfy the boundary conditions $y(-1) = y(1) = 0$. Thus, it is

required to find the extremum of the functional $S(y)$ for the condition $l(y) = \pi$. We form the auxiliary functional

$$I^* = \int_{-1}^{1} (y + \lambda \sqrt{1 + y'^2})\, dx$$

and write for it Euler's equation

$$1 - \lambda \frac{d}{dx} \frac{y'}{\sqrt{1 + y'^2}} = 0.$$

Integrating this equation, we find $\lambda y' / \sqrt{1 + y'^2} = x - c_1$. Solving the last equation with respect to y', we have

$$y' = \pm \frac{x - c_1}{\sqrt{\lambda^2 - (x - c_1)^2}}.$$

Integrating, we obtain

$$y - c_2 = \pm \sqrt{\lambda^2 - (x - c_1)^2}, \text{ or } (x - c_1)^2 + (y - c_2)^2 = \lambda^2.$$

The constants c_1, c_2, and λ are obtained from the boundary conditions and the constraint condition $l = \pi$. We have the system of equations

$$(c_1 - 1)^2 + c_2^2 = \lambda^2,$$
$$(c_1 + 1)^2 + c_2^2 = \lambda^2,$$
$$\lambda \int_{-1}^{1} \frac{dx}{\sqrt{\lambda^2 - (x - c_1)^2}} = \pi.$$

Solving this system, we find: $c_1 = 0$, $c_2 = 0$, and $\lambda = 1$. Thus, the arcs of a circle $y = \sqrt{1 - x^2}$ and $y = -\sqrt{1 - x^2}$ yield the solution of the isoperimetric problem under consideration. The minimum area is equal to $\pi/2$.

Example 3°. Determine the form taken by an absolutely flexible thread of length l suspended in a gravitational field with both ends fixed.

The thread sags so that its potential energy is minimum. The potential energy of the thread is determined with the aid of the integral

$$I(y) = \int_{\gamma} \rho g y\, dl = \int_{x_1}^{x_2} \rho g y \sqrt{1 + y'^2}\, dx,$$

where γ is the curve whose equation is $y = y(x)$, and (x_1, y_1), (x_2, y_2) are the fixed end points of the thread, g is free fall acceleration, and ρ is density. Thus, it is required to find the minimum of the functional $I(y)$ provided that the length of the thread remains constant, that is,

$$l = \int_{x_1}^{x_2} \sqrt{1+y'^2}\, dx = \text{const}.$$

We write the auxiliary functional:

$$I^* = \int_{x_1}^{x_2} (\rho g y \sqrt{1+y'^2} + \lambda \sqrt{1+y'^2})\, dx.$$

The integrand does not contain x explicitly, therefore we write Euler's equation for this functional by formula (7) given in the preceding section:

$$\rho g y + \lambda \sqrt{1+y'^2} - \frac{y'^2(\rho g y+\lambda)}{\sqrt{1+y'^2}} = \frac{1}{c_1}.$$

Let us solve this differential equation. We have:

$$y' = \pm \sqrt{c_1^2(\rho g y+\lambda)^2 - 1};$$

$$\ln\left| c_1(\rho g y+\lambda) + \sqrt{c_1^2(\rho g y+\lambda)^2 - 1}\right| = c_1 x + c_2;$$

$$y = -\frac{\lambda}{\rho g} + c_1 \cosh\left(\frac{x}{c_1} + c_2\right).$$

Thus, we have obtained the equation of the catenary whose shape is taken by the suspended thread.

Let us now compute the length of this catenary:

$$\int_{x_1}^{x_2} \sqrt{1+\sinh^2\left(\frac{x}{c_1}+c_2\right)^2}\, dx = \int_{x_1}^{x_2} \cosh\left(\frac{x}{c_1}+c_2\right) dx$$

$$= c_1\left[\sinh\left(\frac{x_2}{c_1}+c_2\right) - \sinh\left(\frac{x_1}{c_1}+c_2\right)\right].$$

Using the boundary conditions for determining the constants c_1, c_2, and λ, we obtain the system of equations

$$\sinh\left(\frac{x_2}{c_1}+c_2\right)-\sinh\left(\frac{x_1}{c_1}+c_2\right)=\frac{l}{c_1},$$

$$c_1\cosh\left(\frac{x_1}{c_1}+c_2\right)-\frac{\lambda}{g\rho}=y_1,$$

$$c_1\cosh\left(\frac{x_2}{c_1}+c_2\right)-\frac{\lambda}{g\rho}=y_2.$$

The solution of such a system in the general case is carried out approximately, here we are not going to dwell on it at length.

4. The Notion of the Reciprocity Principle. In the preceding subsection, we considered the problem of finding the extremals of functional (1) satisfying the boundary conditions (2) and isoperimetric conditions (13). Closely connected with this problem is the following variation problem. Let I_s be one of the functionals (13), that is,

$$I_s=\int_a^b F_s(x,\ y_1,\ y_1',\ \ldots,\ y_n,\ \ y_n')\,dx. \tag{17}$$

It is required to find the extremals of functional (17) satisfying the boundary conditions (2) and the following isoperimetric conditions:

$$\left.\begin{aligned} I&=\int_a^b F(x,\ y_1,\ y_1',\ \ldots,\ y_n,\ y_n')\,dx,\\ I_j&=\int_a^b F(x,\ y_1,\ y_1',\ \ldots,\ y_n,\ \ y_n')\,dx,\\ &\qquad\qquad j=1,\ \ldots,\ s-1,\quad s+1,\ \ldots,\ m. \end{aligned}\right\} \tag{18}$$

The extremals of such a variation problem are the extremals of the functional

$$\int_a^b\left(F_s+\sum_{j=1,\ j\neq s}^{m}\lambda_jF_j+\lambda_0F_{y_i}\right)dx,$$

that is, they satisfy the system of Euler's equations

$$(F_s)_{y_i} - \frac{d}{dx}(F_s)_{y_i'} + \lambda_0 F_{y_i} - \lambda_0 \frac{d}{dx} F_{y_i'} + \sum_{j=1,\, j\neq s}^{n} \lambda_j \left((F_j)_{y_i} - \frac{d}{dx}(F_j)_{y_i'} \right) = 0, \quad i = 1, 2, \ldots, n.$$

Dividing all the equations of this system by $\lambda_0 \neq 0$, we write it in the form

$$F_{y_i} + \sum_{j=1}^{m} \alpha_j (F_j)_{y_i} - \frac{d}{dx}\left(F_{y_i'} + \sum_{j=1}^{m} \alpha_j (F_j)_{y_i'} \right) = 0, \quad i = 1, 2, \ldots, n,$$

where $\alpha_j = \lambda_j/\lambda_0$, $j = 1, 2, \ldots, s-1, s+1, \ldots, m$, $\lambda_s = 1/\lambda_0$.

Thus, we have a system of Euler's equations for variation problems ((1), (2), (13)). Therefore, the extremals in variation problems ((1), (2), (13)) and in ((17), (18), (2)) are the same. Thus we have established the important property of isoperimetric problems: the variation problems obtained from the problems ((1), (2), (13)) by replacing functional (1) by some functional in conditions (13) have the same extremals. The established property is called the *reciprocity principle*. For instance, the problem on the maximum area bounded by a closed curve of a given length (Example 2) and the problem on the minimum length of a closed curve bounding a given area are reciprocal and have common extremals.

Here, the arcs of a circle $y = \sqrt{1 - x^2}$ and $y = -\sqrt{1 - x^2}$ maximize the functional $\int_0^1 y\, dx$ for the given length $l = \int_0^1 \sqrt{1 + y'^2}\, dx = \pi$. At the same time these curves $y = \pm\sqrt{1 - x^2}$ minimize the functional $\int_0^1 \sqrt{1 + y'^2}\, dx$ for the given area $S = \pi/2$.

CHAPTER 7

Certain Methods of Solving Variation Problems

Sec. 7.1. VARIATION PROBLEMS WITH MOVING BOUNDARIES

1. General Formula of Variation of Functionals. Boundary conditions constitute an integral part of the calculus of variations, and any change in boundary conditions changes the form of the functions minimizing or maximizing a functional. We shall consider variation problems for the functionals defined on those curves whose end points may be displaced in an arbitrary manner, the so-called *variation problems with moving* (or *free*) *boundary points.* Let each curve $y = y(x)$ be defined on its interval $[x_0, x_1]$, generally speaking, different from the interval on which another function is defined. Therefore let us agree about the following: if we consider simultaneously two curves $y = y(x)$ and $y = \tilde{y}(x)$ defined on $[x_0, x_1]$ and $[\tilde{x}_0, \tilde{x}_1]$, respectively, then we extend them in a continuous way, say, along the tangent to the union $[x_0, x_1] \cup [\tilde{x}_0, \tilde{x}_1]$. Shown in Fig. 39 by a broken line is the continuation of the curve $y = y(x)$ onto the interval $[x_1, \tilde{x}_1]$. Let us confine ourselves to a detailed study of a variation problem with moving boundary points of the functional

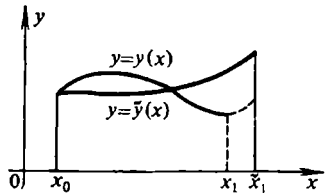


Fig. 39

$$I(y) = \int_{x_0}^{x_1} F(x, y, y')\,dx. \tag{1}$$

We shall assume that the function $F(x, y, z)$ is continuous and has continuous derivatives up to the second order inclusively. The functions $y(x)$ and $y'(x)$ are continuous on the interval under consideration. The same as in the case of two functions $y(x)$ and $\tilde{y}(x)$, we shall regard them continuous on the union $[x_0, x_1] \cup [\tilde{x}_0, \tilde{x}_1]$.

The curves $y(x)$ and $\tilde{y}(x)$ will be regarded close if they are close not only in the sense of the norm of the space $C^{(1)}$ but also in the sense of the proximity of their left-hand end points $P_0(x_0, y(x_0))$, $\tilde{P}_0(\tilde{x}_0, \tilde{y}(\tilde{x}_0))$ and right-hand points $P_1(x_1, y(x_1))$, $\tilde{P}_1(\tilde{x}_1, \tilde{y}(\tilde{x}_1))$. Therefore, for the distance between the two curves $y(x)$ and $\tilde{y}(x)$, we shall take the quantity

$$\rho(y, \tilde{y}) = \|\tilde{y} - y\|_1 + \sqrt{(\tilde{x}_0 - x_0)^2 + (\tilde{y}(\tilde{x}_0) - y(x_0))^2} + \sqrt{(\tilde{x}_1 - x_1)^2 + (\tilde{y}(\tilde{x}_1) - y(x_1))^2}, \qquad (2)$$

where the norm $\|\tilde{y} - y\|_1$ refers to the interval $[x_0, x_1] \cup [\tilde{x}_0, \tilde{x}_1]$. For the sake of simplicity, we shall assume that one end point (on the left) is fixed, while the second end point (on the right) is free (see Fig. 39).

Let us find the increment ΔI of functional (1). Introducing the notation

$$\tilde{x}_1 = x_1 + \delta x_1, \quad \delta y = \tilde{y}(x) - y(x) \quad h(x),$$
$$(\delta y)' = \delta y' = h',$$

we transform the increment:

$$\Delta I = \int_{x_0}^{x_1 + \delta x_1} F(x, y + h, y' + h')\,dx - \int_{x_0}^{x_1} F(x, y, y')\,dx$$
$$\int_{x_0}^{x_1} |F(x, y + h, y' + h') - F(x, y, y')|\,dx$$
$$+ \int_{x_1}^{x_1 + \delta x_1} F(x, y + h, y' + h')\,dx = I_1 + I_2, \qquad (3)$$

where

$$I_1 = \int_{x_0}^{x_1} [F(x, y+h, y'+h') - F(x, y, y')]\,dx,$$

$$I_2 = \int_{x_1}^{x_1+\delta x_1} F(x, y+h, y'+h')\,dx.$$

We first transform the integral I_2 according to the mean-value theorem, and then take advantage of the continuity of the integrand; we get

$$I_2 = F|_{x=x_1+\theta\delta x_1}\delta x_1 = F|_{x=x_1}\delta x_1 + o(\delta x_1).$$

Here, $0 < \theta < 1$, $F|_{x=x_1}$ means the value of the function at the point x_1 and $F|_{x=x_1+\theta\delta x_1}$ means the value at the point $x_1 + \theta\delta x_1$.

Let us now transform the integral I_1, applying Taylor's formula to the integrand:

$$I_1 = \int_{x_0}^{x_1} [F(x, y+h, y'+h') - F(x, y, y')]\,dx$$

$$= \int_{x_0}^{x_1} F_y(x, y, y')\,h\,dx + \int_{x_0}^{x_1} F_{y'}(x, y, y')\,h'\,dx + \alpha(h).$$

It has already been shown that $\alpha(h) = o(\|h\|_1)$, consequently, all the more, $\alpha(h) = o(\rho(\tilde{y}, y))$. We transform the second integral into I_1 by integration by parts:

$$\int_{x_0}^{x_1} F_{y'}h'\,dx = (F_{y'}h)\Big|_{x_0}^{x_1} - \int_{x_0}^{x_1} \frac{d}{dx} F_{y'}h\,dx$$

$$F_{y'}|_{x=x_1}h(x_1) - F_{y'}|_{x=x_0}h(x_0) - \int_{x_0}^{x_1} \frac{d}{dx} F_{y'}h\,dx.$$

The left-hand end point $P_0(x_0, y_0)$ is fixed, therefore $h(x_0) = 0$ and, hence, I_1 can be rewritten as

$$I_1 \int_{x_0}^{x_1} \left(F_y - \frac{d}{dx} F_{y'}\right) h\,dx + F_{y'}|_{x=x_1}h(x_1) + o(\rho(\tilde{y}, y)).$$

The value $h(x_1)$ is the increment of the ordinate at the point P_1 when passing from the curve $y(x)$ to the curve $\tilde{y}(x)$. We denote by δy_1 the increment of the ordinate when passing from the point P_1 to the point $\tilde{P}_1$, i.e.

$$\delta y_1 = \tilde{y}(\tilde{x}_1) - y(x_1) = \tilde{y}(x_1 + \delta x_1) - y_1.$$

It is seen from Fig. 40 that $P_1A = h(x_1)$ and $B\tilde{P}_1 = \delta y_1$. We then express $h(x_1)$ in terms of the increments δx_1 and δy_1:

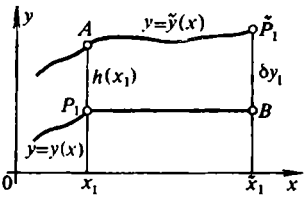


Fig. 40

$$h(x_1) = \tilde{y}(x_1) - y(x_1) = \tilde{y}(x_1) - y_1,$$

and since $y_1 = y(x_1 + \delta x_1) - \delta y_1$,

$$h(x_1) = \tilde{y}(x_1) - \tilde{y}(x_1 + \delta x_1) + \delta y_1.$$

By Lagrange's theorem, $\tilde{y}(x + \delta x_1) - \tilde{y}(x_1) = \tilde{y}'(x_1)\,\delta x_1 + o(\delta x_1)$. The curves $\tilde{y}(x)$ and $y(x)$ are close in the sense of distance (2), therefore

$$\tilde{y}'(x_1) = y'(x_1) + o(\rho(\tilde{y}, y))$$

and

$$h(x_1) = \delta y_1 - y'(x_1)\,\delta x_1 + o(\rho(\tilde{y}, y)).$$

The integral I_1 takes the form

$$I_1 = \int_{x_0}^{x_1} \left(F_y - \frac{d}{dx} F_{y'}\right) h\,dx + F_{y'}|_{x=x_1}\,\delta y_1 - F_{y'}|_{x=x_1} y'(x_1)\,\delta x_1 + o(\rho(\tilde{y}, y)).$$

Substituting I_1 and I_2 into (3), we obtain the expression (for the increment of the functional):

$$\Delta I = \int_{x_0}^{x_1} \left(F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'}\right) h \, \mathrm{d}x + F_{y'}|_{x=x_1} \delta y_1 + (F - y'F_{y'})|_{x=x_1} \delta x_1 + o(\rho(\tilde{y}, y)).$$

Hence it follows that the variation of functional (1), that is, the principal part of the increment ΔI, in the case of one moving boundary point is represented in the form

$$\delta I = \int_{x_0}^{x_1} \left(F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'}\right) h \, \mathrm{d}x + F_{y'}|_{x=x_1} \delta y_1 + (F - y'F_{y'})|_{x=x_1} \delta x_1. \quad (4)$$

And if both end points are movable, then, introducing the notation: $\delta x_0 = \tilde{x}_0 - x_0$, $\delta x_1 = \tilde{x}_1 - x_1$, $\delta y_0 = \tilde{y}(\tilde{x}_0) - y(x_0)$, $\delta y_1 = \tilde{y}(\tilde{x}_1) - y(x_1)$, we write the variation of functional (1) in the form

$$\delta I = \int_{x_0}^{x_1} \left(F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'}\right) h \, \mathrm{d}x + F_{y'}|_{x=x_1} \delta y_1 - F_{y'}|_{x=x_0} \delta y_0 + [(F - y'F_{y'}) \delta x]|_{x_0}^{x_1}, \quad (5)$$

where the expression $\Phi|_{x_0}^{x_1}$ means, as usual, a double substitution, i.e.

$$\Phi|_{x_0}^{x_1} = \Phi(x_1) - \Phi(x_0).$$

Formula (5) is called the *general formula of the variation of a functional* of one function. If $\delta x_0 = \delta x_1 = \delta y_0 = \delta y_1 = 0$, then from (5) we obtain the variation of functional (1) for the case of fixed boundary points.

2. Necessary Condition for an Extremum in a Variation Problem with Moving Boundary Points. In the case of moving boundary points the variation of functional (1) is determined by formula (5), and since for extremal curves the condition

$\delta I = 0$ must be fulfilled, we have

$$\delta I = \int_{x_0}^{x_1} \left(F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'}\right) h \, \mathrm{d}x + F_{y'}|_{x=x_1} \delta y_1 - F_{y'}|_{x=x_0} \delta y_0 + [(F - y'F_{y'}) \delta x]|_{x_0}^{x_1} = 0. \quad (6)$$

If the function $y(x)$ with end points P_0 and P_1 extremizes functional (1) considered on all permissible curves, then, all the more, it will yield an extremum with respect to all the curves having the same end points P_0 and P_1. This means that y extremizes functional (1) on the curves whose end points are fixed, and therefore $y(x)$ satisfies Euler's equation

$$F_y - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'} = 0. \quad (7)$$

Consequently, condition (6) for the extremal curve $y(x)$ turns into the boundary condition

$$F_{y'}|_{x=x_1} \delta y_1 + (F - y'F_{y'})|_{x=x_1} \delta x_1 - F_{y'}|_{x=x_0} \delta y_0 - (F - y'F_{y'})|_{x=x_0} \delta x_0 = 0. \quad (8)$$

From equality (8) we obtain the conditions separately for the left-hand and right-hand end points of the extremal curves. Let us, for instance, analyze the right-hand end point. If the curve $y(x)$ with end points P_0 and P_1 extremizes functional (1) on all permissible curves, it, all the more, will extremize the curves having their left-hand end point at the point P_0; in this case $\delta y_0 = \delta x_0 = 0$, and condition (8) takes the form

$$F_{y'}|_{x=x_1} \delta y_1 + (F - y'F_{y'})|_{x=x_1} \delta x_1 = 0. \quad (9)$$

Analogous reasoning shows that at its left-hand end point the extremal curve $y(x)$ satisfies the condition

$$F_{y'}|_{x=x_0} \delta y_0 + (F - y'F_{y'})|_{x=x_0} \delta x_0 = 0. \quad (10)$$

Thus, if the curve $y = y(x)$ extremizes functional (1) in the case of moving end points, then it satisfies Euler's equation (7) and the boundary conditions (9) and (10).

Conditions (9) and (10) are inconvenient for practical use, since they contain the variations δx_i, δy_i, $i = 1, 2$, and the variation δy_i may depend on δx_i. But we shall use

these conditions in the next two subsections when deriving natural boundary conditions and transversality conditions.

3. Natural Boundary Conditions. In practice, we often encounter the following variation problem: *Find the extremum of functional* (1) *provided that the end points of the permissible curves lie on the straight lines* $x = x_0$ *and* $x = x_1$. In this case $y(x_0)$ and $y(x_1)$ are not specified, but the variations $\delta x_0 = \delta x_1 = 0$. Conditions (9) and (10) are then written in the form

$$F_{y'}|_{x=x_1} \delta y_1 = 0, \quad F_{y'}|_{x=x_0} \delta y_0 = 0.$$

Hence, due to the arbitrariness of δy_0 and δy_1, we get the conditions

$$F_{y'}|_{x=x_1} = 0 \quad \text{and} \quad F_{y'}|_{x=x_0} = 0, \tag{11}$$

which are called the *natural boundary conditions for functional* (1). Thus, in order to find the extremals of functional (1) when the values $y(x_0)$ and $y(x_1)$ are not given, one has to solve Euler's equation $F_y - (\mathrm{d}/\mathrm{d}x) F_{y'} = 0$ and to determine the arbitrary constants from the natural boundary conditions (11).

Example 1°. Find the extremal of the functional

$$I(y) = \int_0^{\pi} (y^2 - y'^2 - 2y \sin x)\, \mathrm{d}x$$

if the left-hand end point is fixed ($y(0) = 0$), and the right-hand end point displaces in the straight line $x = \pi$.

No condition is imposed on the value of the extremal $y(x)$ at the right-hand end point $x = \pi$, therefore, to find the extremal, it is necessary to find the solution of Euler's equation $y'' + y = \sin x$ for the natural boundary condition $F_{y'}|_{x=\pi} = -2y'|_{x=\pi} = 0$. The general solution of Euler's equation is written in the form

$$y = c_1 \cos x + c_2 \sin x - \frac{1}{2} x \cos x.$$

Then from the condition $y(0) = 0$ we find $c_1 = 0$, and from the condition $y'(\pi) = 0$ we obtain the equation

$$y'|_{x=\pi} = \left(c_2 \cos x - \frac{\cos x}{2} + \frac{x \sin x}{2}\right)\Big|_{x=\pi} = -c_2 + \frac{1}{2} = 0,$$

whence $c_2 = 1/2$. Consequently, the desired extremal is the curve $y = (1/2)(\sin x - x \cos x)$.

4. Transversality Condition. Consider the variation problem for the functional

$$I(y) = \int_{x_0}^{x_1} F(x, y, y')\,dx \tag{1}$$

when the end points of the permissible curves move along the given lines $y = \varphi(x)$ and $y = \psi(x)$. For the sake of simplicity, we assume that one end point (say the left-hand one) is fixed, and the other (say the right-hand one) is displaced in the curve $y = \psi(x)$ (Fig. 41). The extremal curve satisfies Euler's equation (7) and the boundary condition (9), that is, the system of equations

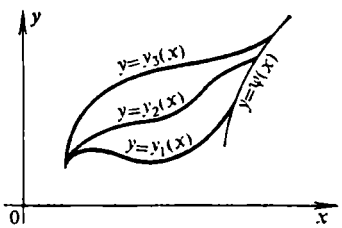


Fig. 41

$$F_y - \frac{d}{dx}F_{y'} = 0, \tag{7}$$

$$F_{y'}|_{x=x_1}\,\delta y_1 + (F - y'F_{y'})|_{x=x_1}\,\delta x_1 = 0. \tag{9}$$

Since the end points of the curves $y(x)$ and $\tilde{y}(x)$ lie on the curve $y = \psi(x)$, we have

$$\delta y_1 = \tilde{y}(x_1 + \delta x_1) - y(x_1) = \psi(x_1 + \delta x_1) - \psi(x_1) = \psi'(x_1)\,\delta x_1 + o(\delta x_1),$$

therefore condition (9) can be written in the form

$$(F - y'F_{y'} + \psi' F_{y'})|_{x=x_1}\,\delta x_1 = 0.$$

Since δx_1 is arbitrary, we have

$$(F - y'F_{y'} + \psi' F_{y'})|_{x=x_1} = 0. \tag{12}$$

Just the same, if the left-hand end point P_0 moves in the curve $y = \varphi(x)$, then the following condition must be fulfilled for it:

$$(F - y'F_{y'} + \varphi' F_{y'})|_{x=x_0} = 0. \tag{13}$$

The boundary conditions (12) and (13) are called the *transversality conditions.*

Thus, when investigating the variation problem with the end points of permissible curves lying on given lines, it is necessary to find the solution of Euler's equation (7) and to determine the two arbitrary constants from the transversality conditions (12) and (13).

The simplest transversality conditions are obtained for the functional of the form

$$I_1(y)=\int_{x_0}^{x_1} f(x,\ y)\sqrt{1+y'^2}\,dx \tag{14}$$

which is frequently encountered in practice. In this case

$$F_{y'}=f(x,\ y)\frac{y'}{\sqrt{1+y'^2}}=\frac{y'F}{1+y'^2}$$

and condition (12) is written in the form

$$F\left(1+\frac{y'(\psi'-y')}{1+y'^2}\right)\Big|_{x=x_1}=\frac{F(1+y'\psi')}{1+y'^2}\Big|_{x=x_1}=0,$$

whence we get the condition for the right-hand end point:

$$y'\psi'|_{x=x_1}=-1.$$

In the same manner, we find the condition for the left-hand end point:

$$y'\varphi'|_{x=x_0}=-1.$$

Thus, for functional (14) the transversality condition is reduced to the condition of orthogonality of the extremals and direction curves $y=\psi(x)$ and $y=\varphi(x)$ at the boundary points.

Example 2°. Find the extremal of the functional

$$I(y)=\int_0^{x_1}\sqrt{1+y}\sqrt{1+y'^2}\,dx$$

if the left-hand end point is fixed ($y(0)=0$), while the right-hand end point moves in the straight line $y+x+1=0$.

Here, the integrand is dependent only on y and y' and is independent of x explicitly, therefore Euler's equation has the form

$$\sqrt{1+y}\sqrt{1+y'^2}-\frac{y'^2\sqrt{1+2}}{\sqrt{1-y'^2}}=c_1,$$

whence we find:

$$y=\frac{(x+c_2)^2}{4c_1^2}+c_1^2-1.$$

The given functional belongs to the functionals of type (14), therefore the transversality condition is reduced to the condition of orthogonality of the desired extremal to the straight line $y=-x-1$. For determining the constants c_1, c_2 and the point of intersection of the extremal and the straight line $y=-x-1$, we have the system of equations:

$$\frac{c_2^2}{4c_1^2}+c_1=1,\quad \frac{x+c_2}{2c_1^2}=1,\quad x+y+1=0,$$
$$y=\frac{(x+c_2)^2}{4c_1^2}+c_1^2-1.$$

Solving this system, we find: $c_1^2=1/5$ and $c_2=4/5$. Consequently, the sought-for extremal is the parabola $y=(5/4)x^2+2x$.

Sec. 7.2. VARIATION PROBLEMS INVOLVING FUNCTIONS OF SEVERAL VARIABLES

1. Variation Problems with Fixed Boundary Values. Let us investigate for an extremum functionals of the functions of several variables. For the sake of simplicity, we will consider the functions of two variables $z=z(x, y)$ which are twice differentiable in a closed plane domain $\bar{G}$.

The increment, or variation, of the argument $\delta z=\tilde{z}(x, y)-z(x, y)$ is also a doubly differentiable function, and the following equalities take place:

$$(\delta z)'_x=\tilde{z}'_x-z'_x=\delta z'_x,\quad (\delta z)'_y=\tilde{z}'_y-z'_y=\delta z'_y.$$

Henceforth, the increment of the argument will be denoted by $h = h(x, y)$, i.e. $h = h(x, y) = \delta z = \tilde{z}(x, y) - z(x, y)$.

By the *variation of the functional*

$$I(z) = \int\int_G F(x, y, z, z'_x, z'_y)\, dx\, dy \qquad (1)$$

we shall understand the principal part of the increment, linear with respect to h, h'_x, h'_y.

Let us consider the variation problem for functional (1), assuming that the domain G does not vary and that the values of the function $z(x, y)$ are given on the boundary γ of the domain G, that is,

$$z|_\gamma = \varphi(x, y). \qquad (2)$$

Geometrically, these conditions mean that a fixed domain G with the boundary γ and a spatial contour Γ whose projection on the xy-plane is γ are given in the xy-plane. All permissible surfaces $z = z(x, y)$ are spanned on this contour Γ. The solution of the variation problem ((1), (2)) within the bounds of the necessary conditions is given by Theorem 1.

For functionals of the functions of several variables there is an analogue of the fundamental lemma of the calculus of variations.

Lemma. *If a fixed function $\alpha(x, y)$ is continuous in a closed domain $\bar{G}$ and if for any function $h(x, y)$ continuous in $\bar{G}$ having in G continuous partial derivatives up to the second order inclusively and vanishing on the boundary γ the integral*

$$\int\int_G \alpha(x, y)\, h(x, y)\, dx\, dy = 0,$$

then $\alpha(x, y) \equiv 0$ in G.

This lemma is proved in the same way as the fundamental lemma of the calculus of variations (see Sec. 6.2), therefore we are not going to dwell on it.

Theorem 1. *If the function $z = z(x, y)$ satisfying conditions* (2) *extremizes functional* (1), *then it is a solution of the Ostrogradsky equation*

$$F_z - \frac{\partial}{\partial x} F_{z'_x} - \frac{\partial}{\partial y} F_{z'_y} = 0. \qquad (3)$$

The solutions of the Ostrogradsky equation are called *extremals.*

Supposing the function F is triply differentiable with respect to its arguments, we can find the variation of functional (1). To do this, we separate the principal part of the increment:

$$\Delta I = \int\int_G [F(x,\ y,\ z+h,\ z'_x+h'_x,\ z'_y+h'_y) - F(x,\ y,\ z,\ z'_x,\ z'_y)]\,dx\,dy.$$

As was the case for a functional of a function of one variable, we find, by transforming the integrand according to Taylor's formula, that the principal part of the increment, which is linear with respect to h, h'_x, h'_y and is the variation of functional (1), has the form

$$\delta I = \int\int_G (F_z h + F_{z'_x} h'_x + F_{z'_y} h'_y)\,dx\,dy. \tag{4}$$

Let us represent this variation in another form. From the formula for differentiating the product

$$\frac{\partial}{\partial x}(F_{z'_x} h) = h\frac{\partial}{\partial x}F_{z'_x} + F_{z'_x} h'_x$$

we have

$$F_{z'_x} h'_x = \frac{\partial}{\partial x}(F_{z'_x} h) - h\frac{\partial}{\partial x}F_{z'_x}.$$

Analogously, we find that

$$F_{z'_y} h'_y = \frac{\partial}{\partial y}(F_{z'_y} h) - h\frac{\partial}{\partial y}F_{z'_y}.$$

Substituting this expression into (4), we obtain

$$\delta I = \int\int_G \left(F_z - \frac{\partial}{\partial x}F_{z'_x} - \frac{\partial}{\partial y}F_{z'_y}\right) h\,dx\,dy + \int\int_G \left(\frac{\partial}{\partial x}(F_{z'_x} h) + \frac{\partial}{\partial y}(F_{z'_y} h)\right) dx\,dy = I_1 + I_2.$$

Let us apply Green's formula to the integral I_2:

$$I_2 = \int\int_G \left(\frac{\partial}{\partial x} (F_{z_x'} h) + \frac{\partial}{\partial y} (F_{z_y'} h) \right) dx\, dy$$

$$= \oint_\gamma F_{z_x'} h\, dy - F_{z_y'} h\, dx.$$

By the hypothesis, all permissible curves attain one and the same value of $\varphi(x, y)$ on the boundary γ, therefore $h(x, y) = 0$ on γ, and, consequently, the integral

$$I_2 = \oint_\gamma -F_{z_y'} h\, dx + F_{z_x'} h\, dy = 0.$$

Thus, the variation δI of functional (1) in the case of fixed boundaries takes the form

$$\delta I = \int\int_G \left(F_z - \frac{\partial}{\partial x} F_{z_x'} - \frac{\partial}{\partial y} F_{z_y'} \right) h\, dx\, dy.$$

By virtue of the necessary condition for the existence of an extremum, the variation of the functional δI is equal to zero. Using the lemma, we obtain that the function $z(x, y)$ giving the extremum of functional (1) satisfies the differential equation (3)*.

Equation (3) is a partial differential equation. Such equations and the relevant methods for their solution are considered in a special course: "Equations of Mathematical Physics". Thus, for instance, the Ostrogradsky equation for the functional

$$I(z) = \int\int_G \left[\left(\frac{\partial z}{\partial x} \right)^2 + \left(\frac{\partial z}{\partial y} \right)^2 + 2zf(x, y) \right] dx\, dy \qquad (5)$$

has the form

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = f(x, y). \qquad (6)$$

* This equation was first obtained by the celebrated Russian mathematician M. Ostrogradsky in 1834.

The last equation is called *Poisson's equation* and is also frequently encountered in problems of mathematical physics.

In solving variation problems, it is required, as a rule, to find the solution of Poisson's equation satisfying the boundary condition $z(x, y)|_\gamma = \varphi(x, y)$, that is, to solve the Dirichlet problem.

Consider a more general case for a functional of the function of n variables.

If $z = z(x_1, x_2, \ldots, x_n)$ is a function of n variables, then in the variation problem for the functional

$$I(z) = \int_\Omega \ldots \int F(x_1, \ldots, x_n, z, z'_{x_1}, \ldots, z'_{x_n})\, dx_1, \ldots, dx_n$$

the necessary condition is given by the Ostrogradsky equation

$$F_z - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_{z'_{x_i}} = 0.$$

In particular, for the functional

$$I(u) = \int_\Omega \int \int \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 + \left(\frac{\partial u}{\partial z} \right)^2 \right] dx\, dy\, dz$$

the Ostrogradsky equation has the form

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0,$$

that is, we obtain Laplace's equation $\Delta u = 0$ for a three dimensional domain. For Laplace's equation the Dirichlet problem is also solved.

2. Natural Boundary Conditions. In the preceding subsection, we investigated functional (1) for an extremum when the domain G did not vary and all permissible curves attained specified values on the boundary γ of the domain G. Let us now consider the variation problem for functional (1) in the case when the domain G does not vary, but the values of the functions $z = z(x, y)$ on the boundary γ are not specified. Geometrically, this means that, instead of one spatial contour Γ, the set of contours Γ_α projected into the plane

contour γ on which the permissible surfaces $z = z(x, y)$ are spanned is considered.

Variation (4) in the case under consideration will have the form

$$\delta I = \int\limits_G \int \left(F_z - \frac{\partial}{\partial x} F_{z'_x} - \frac{\partial}{\partial y} F_{z'_y} \right) h \, dx \, dy + \oint\limits_\gamma - F_{z'_y} h \, dx + F_{z'_x} h \, dy. \qquad (7)$$

In contrast to the variation problem with specified boundary values, in this case $h(x, y)$ does not vanish on the contour γ, therefore the integral $\oint\limits_\gamma - F_{z'_y} h \, dx + F_{z'_x} h \, dy$ does not disappear.

As for the case of the function of one variable, if $z = z(x, y)$ extremizes functional (1) on arbitrary surfaces, then, all the more, it will yield an extremum with respect to all the surfaces spanned on one and the same contour Γ. Therefore $z = z(x, y)$ satisfies the Ostrogradsky equation (3) and from the necessary condition for the existence of an extremum of a functional it follows that for an extremal surface the line integral must be equal to zero, that is,

$$I_1 = \oint\limits_\gamma - F_{z'_y} h \, dx + F_{z'_x} h \, dy = 0.$$

Let the curve γ be represented parametrically: $x = x(t)$, $y = y(t)$, $\alpha \leqslant t \leqslant \beta$. We reduce the line integral I_1 to the definite integral

$$\oint\limits_\gamma - F_{z'_y} h \, dx + F_{z'_x} h \, dy = \int\limits_\alpha^\beta (- F_{z'_y} x' + F_{z'_x} y') \, h \, dt = 0.$$

Since h is arbitrary, from the fundamental lemma of the calculus of variations we conclude that

$$(F_{z'_y} x' - F_{z'_x} y')|_\gamma = 0. \qquad (8)$$

Condition (8) is called the *natural boundary condition for functional* (1).

Let us find, for example, the natural boundary condition for functional (5). We compute the partial derivatives:

$$F_{z'_y} = 2\frac{\partial z}{\partial y}, \quad F_{z'_x} = 2\frac{\partial z}{\partial x}$$

and write condition (8) in the form

$$\left(\frac{\partial z}{\partial y}\frac{dx}{dt} - \frac{\partial z}{\partial x}\frac{dy}{dt}\right)_\gamma = 0. \tag{9}$$

The last relationship can be given another form. Consider the equation of the contour γ in vector form: $\mathbf{r}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j}$. Then

$\mathbf{r}'(t) = x'(t)\,\mathbf{i} + y'(t)\,\mathbf{j}$ is the equation of the tangent,

$\mathbf{n}(t) = y'(t)\,\mathbf{i} - x'(t)\,\mathbf{j}$ is the equation of the normal.

Let us now compute the normal derivative of $z(x, y)$:

$$\frac{dz}{d\mathbf{n}} = \operatorname{grad} z \cdot \mathbf{n}^0 = \left(\frac{\partial z}{\partial x}\mathbf{i} + \frac{\partial z}{\partial y}\mathbf{j}\right) \cdot \frac{y'\mathbf{i} - x'\mathbf{j}}{\sqrt{x'^2 + y'^2}}$$

$$= \frac{1}{\sqrt{x'^2 + y'^2}}\left(\frac{\partial z}{\partial x}\frac{dy}{dt} - \frac{\partial z}{\partial y}\frac{dx}{dt}\right).$$

Using equality (9), we conclude that the natural boundary condition for functional (5) consists in that on the curve γ the normal derivative of the extremal surface is equal to zero:

$$\left.\frac{dz}{d\mathbf{n}}\right|_\gamma = 0.$$

But this means that finding the extremal of functional (5) with natural boundary condition is reduced to solving the Neumann problem for Poisson's equation.

Sec. 7.3. CONNECTION OF VARIATION PROBLEMS WITH DIFFERENTIAL EQUATIONS

1. Deriving the Equations of Vibrations of a String and a Membrane. Let us apply the results obtained in the preceding section to derive the equations of vibrations of a string and a membrane. We shall rely here on the principle of least action, formulated by M. Ostrogradsky and W. Hamilton, which is the leading variation principle in mechanics. This principle states that among all possible motions

of a system of material points there is a motion which minimizes the functional

$$I = \int_{t_0}^{t} (T - U)\, dt, \tag{1}$$

where T is the kinetic and U the potential energy of the system. Functional (1) is called the *action*.

Let us first use the Ostrogradsky-Hamilton principle to derive the equation for a vibrating string. Consider the motion of a string (a flexible material thread of length l) with line density $\rho = \text{const}$. Let in the equilibrium position the string be directed along the x-axis. We denote by $u(x, t)$ the displacement of the string from the equilibrium position at the point x at the instant of time t. We shall consider only transverse vibrations of the string, assuming that the motion occurs in one plane and that all points of the string move perpendicularly to the x-axis. Suppose that the end points of the string $x = 0$ and $x = l$ are fixed, that is, $u(0, t) = u(l, t) = 0$. The kinetic energy of the string is determined by the relationship

$$T = \int_0^l \frac{\rho}{2} u_t'^2(x, t)\, dx. \tag{2}$$

Let us find the expression for the potential energy of the string. The potential energy is equal to the work which must be done to bring the string from the equilibrium position to the position under consideration. Since the string is regarded to be absolutely flexible, the entire work is spent on its stretching (but not on its bending) and overcoming the external forces. Let the string tension be maintained at a constant level: $k = \text{const}$. We shall confine ourselves to considering small-amplitude vibrations of the string, that is, we shall hold that the displacement $u(x, t)$ and also the derivative $\partial u/\partial x$ are so small that we may neglect the terms containing u and $\partial u/\partial x$ raised to powers higher than two. Let us consider an element of the string in two positions: initial and final. The work ΔU_1 spent on stretching the element Δx (Fig. 42) is equal to the product of the ten-

sion k by the magnitude of stretching $\sqrt{1+u_x'^2}\,\Delta x - \Delta x$, that is,

$$\Delta U_1 = k\left(\sqrt{1+u_x'^2} - 1\right)\Delta x.$$

Applying Taylor's formula and rejecting the terms of higher order of smallness than $u_x'^2\,\Delta x$, we have

$$\Delta U_1 = k\left(\sqrt{1+u_x'^2} - 1\right)\Delta x \approx \frac{1}{2} k u_x'^2\,\Delta x.$$

Then for the whole string this work is computed with the aid of the integral

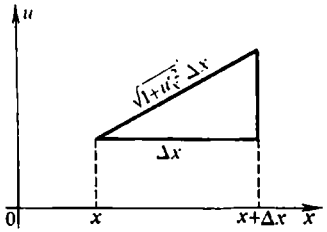


Fig. 42

$$U_1 = \int_0^l \frac{k}{2} u_x'^2\,dx.$$

Let us now assume that the string is also acted upon by an external restoring force $f(x, t)$ perpendicular to the string at the moment of its equilibrium position and calculated per unit mass. The external forces acting on the string element Δx do the work equal to the product of the force $\rho f\,\Delta x$ by the path $u(x, t)$, i.e. $\Delta U_2 = \rho u f\,\Delta x$, and the total work of the external forces is determined by the relationship

$$U_2 = \int_0^l \rho u f\,dx.$$

The potential energy of the whole string at the instant of time t equals the difference between the works U_1 and U_2:

$$U = U_1 - U_2 = \int_0^l \left(\frac{k}{2} u_x'^2 - \rho u f\right) dx.$$

Let us now write for the case under consideration the action during the interval $[t_0, t_1]$ determined by functional (1):

$$I(u) = \int_{t_0}^{t_1}\int_0^l \left(\frac{\rho}{2} u_t'^2 - \frac{k}{2} T_0 u_x'^2 + \rho u f\right) dx\,dt.$$

According to the principle of least action, this functional reaches a minimum on the function $u(x, t)$ satisfying the Ostrogradsky equation (see (3) in the preceding section) which for our functional has the form

$$\rho f - \frac{\partial}{\partial t}(\rho u_t') + \frac{\partial}{\partial x}(k u_x') = 0.$$

Putting $a^2 = k/\rho$, we obtain the equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x, t) \tag{3}$$

called the *equation of string vibrations*. If the external forces are absent, i.e. $f(x, t) = 0$, then we obtain the equation of free vibrations of an elastic string

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}. \tag{4}$$

Just in a similar way, we can obtain the equation of membrane vibrations. Let us denote by $u(x, y, t)$ the deviation of the point (x, y) of the membrane from the equilibrium position at time t, and by $f(x, y, t)$ the external force perpendicular to the membrane in the equilibrium position and calculated per unit mass. Let the density ρ of the membrane and tension k be constant. The kinetic energy of the membrane at instant t is computed by the formula

$$T = \int\int_G \frac{\rho}{2} u_t'^2 \, dx \, dy,$$

and the work spent on deforming the membrane element is equal to

$$k\sqrt{1 + u_x'^2 + u_y'^2}\,\Delta x\,\Delta y - k\,\Delta x\,\Delta y \approx \frac{1}{2} k (u_x'^2 + u_y'^2)\,\Delta x\,\Delta y.$$

The potential energy of the entire membrane is equal to the difference between the work U_1 spent on deforming the membrane and the work U_2 done by the external forces, that is,

$$U = U_1 - U_2 = \int\int_G \left(\frac{k}{2}(u_x'^2 + u_y'^2) - \rho u f\right) dx\, dy.$$

The action during the time interval $[t_0, t_1]$ determined by functional (1) is represented in the form

$$I(u) = \int_{t_0}^{t_1} \iint_G \left(\frac{\rho}{2} u_t'^2 - \frac{k}{2} (u_x'^2 + u_y'^2) + \rho u f \right) dx\, dy\, dt.$$

According to the principle of least action, the function $u(x, y, t)$ describing the real motion of the membrane must minimize the functional $I(u)$. Thus, the *membrane motion equation* is obtained from the Ostrogradsky equation and is written in the form

$$\frac{\partial^2 u}{\partial t^2} = a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + f(x, y, t), \tag{5}$$

where $a^2 = k/\rho$. In particular, for free vibrations of the membrane we have the equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right). \tag{6}$$

2. Variation Problems Connected with Poisson's Equation. From the aforegoing it is seen that solving variation problems is reduced to solving boundary-value problems for ordinary differential equation (for a functional of the functions of one variable) and for partial differential equations (for a functional of the functions of several variables). It turns out that solving certain boundary-value problems, which play an important role in applications, is equivalent to solving a concrete variation problem. Let us establish the relation between variation problems and boundary-value problems for Poisson's equation. In Sec. 7.2, it was shown that the problem involving the minimum of the functional

$$I(z) = \iint_G \left[\left(\frac{\partial z}{\partial x} \right)^2 + \left(\frac{\partial z}{\partial y} \right)^2 + 2zf(x, y) \right] dx\, dy \tag{7}$$

provided that on the boundary γ of the domain G the permissible functions $z = z(x, y)$ attain the given values

$$z(x, y)|_\gamma = \varphi(P), \quad P(x, y) \in \gamma \tag{8}$$

is reduced to solving Poisson's equation

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = f(x, y). \tag{9}$$

We will assume that the function $f(x, y)$ is continuous in the closed domain $\bar{G}$ and that $\varphi(P)$ is continuous on γ.

Theorem 1. *The problem of solving equation* (9) *satisfying the boundary condition* (8) *is equivalent to the variation problem of finding the minimum of functional* (7) *under the same condition* (8).

Reduction of the variation problem (7) with condition (8) to the boundary-value problem (9) with the same condition was performed in the preceding section. Let us prove the converse. Let $z = z_0(x, y)$ be the solution of the boundary-value problem ((9), (8)). Let us show that the solution $z_0(x, y)$ gives an absolute minimum to functional (8). To this end, let us compute the increment of the functional $I(z)$. Since for the variation of the function $z = z(x, y)$ we have the relationships

$$(\delta z)'_x = \delta z'_x = h'_x \quad \text{and} \quad (\delta z)'_y = \delta z'_y = h'_y,$$

for the increment of the functional we obtain

$$\Delta I = \iint\limits_G \left[\left(\frac{\partial z_0}{\partial x} + \frac{\partial h}{\partial x} \right)^2 + \left(\frac{\partial z_0}{\partial y} + \frac{\partial h}{\partial y} \right)^2 + 2f(z_0 + h) - \left(\frac{\partial z_0}{\partial x} \right)^2 - \left(\frac{\partial z_0}{\partial y} \right)^2 - 2fz_0 \right] dx\, dy$$

$$= 2 \iint\limits_G \left(\frac{\partial z_0}{\partial x} \frac{\partial h}{\partial x} + \frac{\partial z_0}{\partial y} \frac{\partial h}{\partial y} + fh \right) dx\, dy$$

$$+ \iint\limits_G \left[\left(\frac{\partial h}{\partial x} \right)^2 + \left(\frac{\partial h}{\partial y} \right)^2 \right] dx\, dy = 2I_1 + I_2. \qquad (10)$$

Since all permissible functions satisfy the boundary condition (8), $h(x, y) \not\equiv \text{const}$ in the domain G, therefore

$$I_2 = \iint\limits_G \left[\left(\frac{\partial h}{\partial x} \right)^2 + \left(\frac{\partial h}{\partial y} \right)^2 \right] dx\, dy > 0.$$

Let us transform the first integral I_1 and show that it equals zero; we have

$$I_1 = \iint\limits_G \left(\frac{\partial z_0}{\partial x} \frac{\partial h}{\partial x} + h \frac{\partial^2 z_0}{\partial x^2} + \frac{\partial z_0}{\partial y} \frac{\partial h}{\partial y} + h \frac{\partial^2 z_0}{\partial y^2} \right) dx\, dy$$

$$-\iint\limits_G \left(\frac{\partial^2 z_0}{\partial x^2}+\frac{\partial^2 z_0}{\partial y^2}-f\right) h\,dx\,dy$$

$$=\iint\limits_G \left[\frac{\partial}{\partial x}\left(h\frac{\partial z_0}{\partial x}\right)+\frac{\partial}{\partial y}\left(h\frac{\partial z_0}{\partial y}\right)\right] dx\,dy$$

$$-\iint\limits_G \left(\frac{\partial^2 z_0}{\partial x^2}+\frac{\partial^2 z_0}{\partial y^2}-f\right) h\,dx\,dy = I_3 + I_4.$$

The function $z_0(x, y)$ is a solution of equation (9) therefore

$$I_4 = \iint\limits_G \left(\frac{\partial^2 z_0}{\partial x^2}+\frac{\partial^2 z_0}{\partial y^2}-f\right) h\,dx\,dy = 0.$$

Applying Green's formula to the integral I_3 and bearing in mind that $h|_\gamma = 0$, we get

$$I_3 = \iint\limits_G \left[\frac{\partial}{\partial x}\left(h\frac{\partial z_0}{\partial x}\right)+\frac{\partial}{\partial y}\left(h\frac{\partial z_0}{\partial y}\right)\right] dx\,dy$$

$$= \oint\limits_\gamma -h\frac{\partial z_0}{\partial y}\,dx + h\frac{\partial z_0}{\partial x}\,dy = 0.$$

Thus, the increment $\Delta I = I_3 + I_4 + I_2 = I_2 > 0$, and, consequently, the function $z_0(x, y)$ minimizes functional (7).

The most general boundary-value problem for Poisson's equation (9) is the so-called *third boundary-value problem*, when the boundary conditions have the form

$$\left(\frac{dz}{dn}+\alpha z\right)\Big|_\gamma = \varphi(P). \tag{11}$$

It is possible to show that the problem concerning the minimum of the functional

$$I(z) = \iint\limits_G \left[\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2+2fz\right] dx\,dy$$

$$+\oint\limits_\gamma (\alpha z^2+2\varphi z)\,dl, \tag{12}$$

when no boundary conditions are imposed on the function $z = z(x, y)$, leads to the third boundary-value problem for Poisson's equation.

Functional (12) can be given the following mechanical interpretation. The integral

$$\iint_G \left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 2fz\right] dx\, dy$$

(with an accuracy to a constant multiplier) represents the potential energy of a stretched membrane (see Sec. 7.3). The line integral

$$\oint_\gamma (\alpha z^2 + 2\varphi z)\, dl$$

is added to the potential energy of the membrane if its contour is movable and is acted upon by an external force with line density $\varphi(P)$ and elastic forces tending to keep the boundary in the equilibrium position whose modulus of elasticity per unit length is equal to $\alpha(P)$.

Sec. 7.4. DIRECT METHODS IN THE CALCULUS OF VARIATIONS

1. Direct Methods. Approximate methods yielding a direct solution of variation problems are called the *direct methods of the calculus of variations*. The most noted among them are the methods proposed by W. Ritz, B. Galerkin, and L. Kantorovich. The development of the direct methods makes it possible to find the solutions of variation problems with any degree of accuracy. These methods turned out to be useful not only for direct solving variation problems but they also have been widely applied in other fields of mathematics, in particular, in solving boundary-value problems for differential equations.

The use of the direct methods is based on the following idea. Consider, for the sake of definiteness, the problem of finding the minimum of the functional $I(y)$ defined on a certain set M of permissible curves. In order for this problem to have sense, it should be assumed that in class M there are curves for which $I(y) < \infty$ and besides, $\inf I(y) = \mu > -\infty$. These restrictions will make it possible to construct approximately the curve on which the functional $I(y)$ reaches a minimum.

The basis of the direct methods is the concept of a minimizing sequence.

The sequence $y_1, \ldots, y_n, \ldots$ is said to be *minimizing* if the following limit relation holds true:

$$\lim_{n\to\infty} I(y_n) = \mu. \tag{1}$$

If on the considered set M of permissible curves the conditions $\inf I(y) = \mu > -\infty$ and $\exists\, y \in M,\ I(y) < \infty$ are fulfilled, then, by virtue of the definition of the greatest lower bound, a minimizing sequence is existent. Suppose that for the minimizing sequence $y_1, y_2, \ldots, y_n, \ldots$ there exists a limit curve $y^* \in M$, and if the passage to the limit

$$\lim_{n\to\infty} I(y_n) = I(\lim_{n\to\infty} y_n) = I(y^*) \tag{2}$$

turns to be valid, then $I(y^*) = \mu$, that is, the limit curve $y^* \in M$ will be the solution of the problem under consideration.

Thus, the solution of a variation problem by a direct method is made up of: (1) constructing a minimizing sequence $y_1, \ldots, y_n, \ldots$, (2) proving that this sequence has a limit curve $y^* \in M$, and (3) proving the validity of the limit passage (2).

The terms of a minimizing sequence may be regarded as approximate solutions of an appropriate variation problem.

The foundation of the direct methods is the construction of the minimizing sequence of functions which is always possible if only $\inf I(y) > -\infty$. Each of the direct methods used in the calculus of variations is characterized just by the way in which minimizing sequences are constructed. But it should be noted that although a minimizing sequence can be constructed in any variation problem, the limit curve of such a sequence may not exist. The question of the existence of the limit curve is rather complicated. It is solved for a broad class of variation problems, but we are not going to dwell on this.

2. The Ritz Method. One of the methods of constructing a minimizing sequence was suggested by W. Ritz in 1908.*

* The theoretical proof of this method was given later, in particular, in the works by the Soviet mathematicians N. Krylov and N. Bogolyubov in 1929.

Let there be sought a minimum of the functional $I(y)$ defined on a certain set lying in a normed linear space E. We shall assume that the permissible functions $y = y(P)$ satisfy homogeneous boundary conditions. This means the following: if $y(P)$ is a function of two variables defined in the domain G with the boundary γ, then on γ the function $y(P)$ satisfies one of the conditions:

$$\text{(a) } y(P)|_{P\in\gamma}=0, \quad \text{(b) } \frac{dy}{dn}\Big|_{P\in\gamma}=0, \quad \text{(c) } \left(y+\alpha\frac{dy}{dn}\right)_{P\in\gamma}=0.$$

The set of permissible functions $y(P) \in E$ satisfying homogeneous boundary conditions forms in R a subspace which will be denoted by M. Let us choose a sequence of functions

$$\varphi_1(P),\ \varphi_2(P),\ \ldots,\ \varphi_n(P),\ \ldots \tag{3}$$

satisfying the following three conditions:

(1°) all $\varphi_n(P) \in M$;

(2°) for any n the totality of functions $\varphi_1(P), \ldots, \varphi_n(P)$ is linearly independent;

(3°) sequence (3) is complete, that is, $\forall\, y \in M$ and $\forall\, \delta > 0$ there is a linear combination

$$y_n = \sum_{k=1}^{n} c_k \varphi_k(P) \tag{4}$$

such that $\| y - y_n \| < \delta$.

Sequence (3) satisfying these conditions is referred to as a *sequence of coordinate functions.*

Let us substitute the linear combination (4) into the functional $I(y)$, thus turning the functional $I(y)$ into a function of n variables $c_1, c_2, \ldots, c_n$:

$$I(y_n) = I\left(\sum_{k=1}^{n} c_k \varphi_k\right) = \Phi(c_1, \ldots, c_n) \tag{5}$$

and let us find the coefficients $c_1, \ldots, c_n$ so that functional (5) attains minimum values.

Thus, the problem of finding the minimum of the functional $I(y)$ is reduced to the problem of finding the minimum of the function $\Phi(c_1, \ldots, c_n)$ of n variables which is by far simpler than the problem of finding the minimum of a functional.

From the necessary condition for a minimum of the function $\Phi(c_1, \ldots, c_n)$ we obtain the system of equations for determining the coefficients $c_1, \ldots, c_n$:

$$\frac{\partial \Phi}{\partial c_k} = 0, \quad k = 1, 2, \ldots, n.$$

Substituting the coefficients c_k found from this system into function (4), we obtain an approximation for the extremal of the functional $I(y)$ by the Ritz method.

Let us give the conditions of the existence for the obtained minimizing sequence. Let us denote by μ_n the minimum of the function $\Phi(c_1, \ldots, c_n)$:

$$\mu_n = \min_{c_i} \Phi(c_1, \ldots, c_n) = \min_{c_i} I\left(\sum_{k=1}^{n} c_k \varphi_k\right).$$

For each $n = 1, 2, \ldots$ we construct the approximation $y_n(P)$ and find the minimum μ_n; then we obtain the functional sequence of approximations

$$y_1(P), \; y_2(P), \; \ldots, \; y_n(P), \; \ldots$$

and the number sequence

$$\mu_1, \; \mu_2, \; \ldots, \; \mu_n, \; \ldots.$$

Assuming the existence of the curve y^* which realizes the minimum of the functional $I(y)$, let us consider the question concerning the conditions under which it is possible to assert that the constructed sequence $y_1, y_2, \ldots, y_n, \ldots$ will be minimizing, that is, $\lim_{n \to \infty} \mu_n = \mu$, where μ is the minimum of the functional $I(y)$.

The following theorem gives the answer to this question.

Theorem 1. *If the functional $I(y)$ is continuous in the space E, and the system of functions* (3) *satisfies conditions* (1°), (2°), *and* (3°), *then the sequence $y_1, y_2, \ldots, y_n, \ldots$ is minimizing.*

It is required to prove that $\lim_{n \to \infty} \mu_n = \mu$. The sequence $\mu_1, \mu_2, \ldots, \mu_n$ is non-increasing, that is, $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_n \geqslant \mu_{n+1} \geqslant \ldots$. This assertion follows from the fact that the linear combinations $y_{n+1} = \sum_{k=1}^{n+1} c_k \varphi_k(x)$

contain all linear combinations $y_n = \sum_{k=1}^{n} c_k \varphi_k(x)$, therefore

$$\mu_n = \min_{c_i} I\left[\sum_{k=1}^{n} c_k \varphi_k(x)\right] \geqslant \min_{c_i} I\left[\sum_{k=1}^{n+1} c_k \varphi_k(x)\right] = \mu_{n+1}.$$

Let y^* be the curve on which the minimum of the functional $I(y)$ is realized, that is,

$$\min_{y} I(y) = I(y^*) = \mu.$$

The continuity of the functional implies that $\forall\, \varepsilon > 0\ \exists\, \delta > 0$ such that

$$|I(y) - I(y^*)| = |I(y) - \mu| < \varepsilon \tag{6}$$

as soon as $\| y - y^* \| < \delta$.

The system of functions (3) is complete, therefore among the linear combinations of form (4) there is $\tilde{y}_{n_0}$ such that $\| \tilde{y}_{n_0} - y^* \| < \delta$. Inequality (6) is fulfilled for any function y satisfying the condition $\| y - y^* \| < \delta$, in particular, also for $\tilde{y}_{n_0}$:

$$|I(\tilde{y}_{n_0}) - \mu| < \varepsilon, \quad \text{or} \quad \mu \leqslant I(\tilde{y}_{n_0}) < \mu + \varepsilon.$$

From the equality $\mu_{n_0} = \min I(\tilde{y}_{n_0})$ it follows that $\mu_{n_0} \leqslant I(\tilde{y}_{n_0})$. Taking into consideration the above inequality, we get

$$\mu \leqslant \mu_{n_0} \leqslant I(\tilde{y}_{n_0}) < \mu + \varepsilon,$$

the sequence $\mu_1, \ldots, \mu_n, \ldots$ is non-increasing, therefore the inequality $\mu < \mu_n < \mu + \varepsilon$ will be fulfilled for any $n \geqslant n_0$.

Consequently, there exists the limit $\lim_{n\to\infty} \mu_n = \mu$.

Since the problem of estimating the error of approximation by the Ritz method is rather complicated, we are not going to consider the accuracy of the obtained results. The only point we should like to indicate is that the rapidity of convergence of the approximations for a given variation problem depends both on the problem under consideration and on the choice of the functions $\varphi_1(P), \varphi_2(P), \ldots$

$\ldots, \varphi_n(P), \ldots$. The choice of the sequence of functions $\varphi_1, \ldots, \varphi_n, \ldots$ considerably affects the degree of complication of further computations and the accuracy of the result obtained. The proper choice ensures a successful application of this method. It should be underlined that in many cases it suffices to take a linear combination of quite a small number of functions $\varphi_n(P)$ (four, three, and sometimes even fewer) in order to obtain a quite satisfactory approximation to the exact solution. Let us apply the Ritz method to solving variation problems for functionals of one and two variables in the general form.

Example 1°. Making use of the Ritz method, construct the nth approximation for the function minimizing the functional

$$I(y) = \int_a^b F(x, y, y')\,dx \tag{7}$$

defined on a certain set of the space $C^{(1)}$.

Functional (7) is continuous in the space $C^{(1)}$. Let us suppose that there exists a curve y^* realizing the minimum of this functional. If the system of functions $\varphi_1(x), \ldots, \varphi_n(x), \ldots$ satisfies the requirements of Theorem 1, then the sequence $y_1, \ldots, y_n, \ldots$ constructed by the Ritz method will be minimizing. In practice, the sequence of coordinate functions $\varphi_1(x), \ldots, \varphi_n(x), \ldots$ is frequently formed with the aid of the sequences $1, x, x^2, \ldots, x^n, \ldots$ or $\sin x, \sin 2x, \ldots, \sin nx, \ldots$. In particular, if the permissible curves of functional (7) satisfy the homogeneous conditions $y(a) = y(b) = 0$, then for coordinate functions $\varphi_n(x)$ we can take, for instance,

$$\varphi_n(x) = (x-a)(b-x)x^n, \quad n = 0, 1, \ldots, \text{ or}$$

$$\varphi_n(x) = \sin\frac{\pi n(x-a)}{b-a}, \quad n = 1, 2, \ldots.$$

If the conditions are nonhomogeneous, say $y(a) = A$ and $y(b) = B$, then it is simplest to seek the solution of the variation problem in the form

$$y_n = \varphi_0(x) + \sum_{k=1}^{n} c_k \varphi_k(x),$$

where $\varphi_0(x)$ satisfies the given boundary conditions $\varphi_0(a) = A$ and $\varphi_0(b) = B$, the rest of $\varphi_k(x)$ satisfying the homogeneous boundary conditions $\varphi_k(a) = \varphi_k(b) = 0$.

For the function $\varphi_0(x)$ we can choose, for instance, the linear function

$$\varphi_0(x) = \frac{B-A}{b-a}(x-a) + A.$$

According to the Ritz method, we seek the solution of the variation problem in the form

$$y_n = \sum_{k=1}^{n} c_k \varphi_k(x).$$

Substituting y_n into functional (7), we get the function of n variables $c_1, \ldots, c_n$

$$\Phi(c_1, \ldots, c_n) = \int_a^b F\left(x, \sum_{k=1}^{n} c_k \varphi_k(x), \sum_{k=1}^{n} c_k \varphi_k'(x)\right) dx,$$

which should be investigated for an extremum, the coefficients $c_1, \ldots, c_n$ are found from the system

$$\frac{\partial \Phi}{\partial c_k} = 0, \quad k = 1, 2, \ldots, n.$$

Solving these systems is, generally speaking, a very complicated problem. It is considerably simplified if the functional $I(y)$ quadratic with respect to the unknown function and its derivatives is investigated for an extremum. In this case the system of equations will be linear with respect to the coefficients c_k.

The Ritz method is also applicable to functionals of a function of several variables.

Example 2°. Find the extremum of the functional

$$I(z) = \iint_G F(x, y, z, z_x', z_y')\, dx\, dy$$

if the permissible curves satisfy the boundary conditions $z(x, y)|_\gamma = 0$.

We shall seek an approximate solution of the extremal in the form

$$z_n = \sum_{k=1}^{n} c_k \varphi_k (x, y),$$

where all $\varphi_k (x, y)|_\gamma = 0$.

The sequence of coordinate functions $\varphi_1, \ldots, \varphi_n$ can be constructed in the following way: we find the function $\omega = \omega (x, y)$ continuous together with partial derivatives in the domain $\bar{G}$ and satisfying the conditions

$$\omega (x, y) > 0 \quad \text{in } G \quad \text{and} \quad \omega (x, y)|_\gamma = 0.$$

Then the sequence of coordinate functions will be represented by the system of functions

$$\varphi_1 = \omega, \quad \varphi_2 = \omega x, \quad \varphi_3 = \omega y, \quad \varphi_4 = \omega x^2, \quad \varphi_5 = \omega y^2, \ldots .$$

The function $\omega = \omega (x, y)$ can be chosen as follows: if the contour γ has the equation $\eta (x, y) = 0$, we set $\omega (x, y) = \pm\eta (x, y)$, where the sign is chosen so as to fulfill the condition $\omega (x, y) > 0$. If, for instance, γ is a circle $x^2 + y^2 = R^2$, then $\omega^2 = R^2 - x^2 - y^2$; if γ is the contour of a rectangle $a \leqslant x \leqslant b, c \leqslant y \leqslant d$, then $\omega (x, y) = (b - x)(x - a)(d - y)(y - c)$. When solving variation problems by the Ritz method, it is very appropriate to make use of digit computers.

3. **Galerkin's Method.** In 1915 B. Galerkin suggested a more general and universal method of the solution of boundary-value problems. One of the advantages of Galerkin's method, which is now widely used, consists in that it can be directly applied to the solution of boundary-value problems of both ordinary differential equations and partial differential equations without reducing them first to variation problems. This means that if a certain boundary-value problem is solved for the differential equation $L(z) = 0$ (where L is a certain differential operator) by the Ritz method, then it should be first reduced to a variation problem. In Galerkin's method such a reduction is not required. Besides, Galerkin's method is applicable to a broader class of problems and is simpler in use.

Let us dwell on the underlying idea of Galerkin's method. Let an unknown function $z (P)$ satisfy in a domain G a differ-

ential equation (ordinary or partial)

$$L(z(P)) = 0 \tag{8}$$

and certain homogeneous boundary conditions. We then choose a sequence of coordinate functions $\varphi_1(P)$, $\varphi_2(P)$, ... , $\varphi_n(P)$, ... As always, we suppose that all functions $\varphi_n(P)$ are continuously differentiable a needed number of times in the closed domain $\bar{G} = G + \gamma$ and on the boundary γ satisfy homogeneous conditions. Then the function $z_n(P) = \sum_{k=1}^{n} c_k \varphi_k(P)$ satisfies the same boundary conditions for arbitrary constants c_k. The coefficients c_k are determined from the condition that the left-hand side of equation (8) becomes orthogonal to the functions $\varphi_1(P)$, $\varphi_2(P)$, ... , $\varphi_n(P)$ upon replacing $z(P)$ by the functions $z_n(P)$.

Let us give some considerations concerning the justification of finding the coefficients c_k. The demand that the function $\tilde{z}(P)$ representable in the form of a series in coordinate functions $\tilde{z}(P) = \sum_{k=1}^{\infty} c_k \varphi_k(P)$ be a solution of equation (8), i.e. $L(\tilde{z}(P)) \equiv 0$, is equivalent to the requirement of the orthogonality of the expression $L(\tilde{z})$ to all functions of the system $\varphi_1(P)$, . . . , $\varphi_n(P)$, ... But having at our disposal only n functions $\varphi_k(x)$, $k = 1, 2, \ldots, n$, we can find only n constants from the condition of the orthogonality of $L(z_n(P))$ to the functions $\varphi_k(x)$, $k = 1, 2, \ldots, n$.

For variation problems, Galerkin's method is closely interrelated with the Ritz method. The variation problem for the functional

$$I(y) = \int_a^b F(x, y, y')\,dx \tag{9}$$

with the given boundary conditions $y(a) = y(b) = 0$ is reduced to the boundary-value problem for the equation

$$L(y) = F_y - \frac{d}{dx} F_{y'} = 0 \tag{10}$$

with the same conditions: $y(a) = y(b) = 0$.

Let $\varphi_1(x), \ldots, \varphi_n(x), \ldots$ be a coordinate system of functions. We will show that if the solution of the variation problem (9) is sought in the form

$$y_n = \sum_{k=1}^{n} c_k \varphi_k(x), \tag{11}$$

determining the coefficients by the Ritz method, and the solution of the boundary-value problem (10) is also sought in form (11), finding the coefficients by Galerkin's method, then we obtain two equivalent systems for determining the coefficients c_k. Putting y_n into functional (9), we get

$$I(y_n) = \int_a^b F\left(x, \sum_{k=1}^{n} c_k \varphi_k, \sum_{k=1}^{n} c_k \varphi_k'\right) dx = \Phi(c_1, \ldots, c_n).$$

We then find the partial derivatives

$$\frac{\partial \Phi}{\partial c_k} = \int_a^b (F_y \varphi_k + F_{y'} \varphi_k') \, dx, \quad k = 1, 2, \ldots, n,$$

and transform them by integrating by parts

$$\frac{\partial \Phi}{\partial c_k} = \int_a^b \left(F_y - \frac{d}{dx} F_{y'}\right) \varphi_k \, dx + F_{y'} \varphi_k' \Big|_a^b.$$

But for $\varphi_k(x)$ the conditions $\varphi_k(a) = \varphi_k(b) = 0$ are satisfied and therefore

$$\frac{\partial \Phi}{\partial c_k} = \int_a^b \left(F_y - \frac{d}{dx} F_{y'}\right) \varphi_k \, dx.$$

When applying the Ritz method, the coefficients c_k are found from the system

$$\frac{\partial \Phi}{\partial c_k} = 0, \quad k = 1, \ldots, n,$$

which has been reduced to the form

$$\int_a^b \left(F_y - \frac{d}{dx} F_{y'}\right) \varphi_k \, dx = 0, \quad k = 1, 2, \ldots, n.$$

The last equalities mean that the function

$$L(y_n) = F_y - \frac{d}{dx} F_{y'}$$

is orthogonal to the functions $\varphi_1(x), \ldots, \varphi_n(x)$ on the interval $[a, b]$. Thus, the Ritz method and Galerkin's method lead to the equivalent systems for determining the coefficients c_k. But Galerkin's method involves simpler calculations.

Example 3°. Find the minimum of the functional

$$I(y) = \int_0^1 (y'^2 e^{x^3} + y^2 e^x + xy)\, dx \qquad (12)$$

for the given boundary conditions: $y(0) = y(1) = 0$.

This problem is equivalent to the boundary-value problem for the differential equation

$$L(y) = 2e^{x^3}(y'' + 2xy') - 2ye^x - x = 0 \qquad (13)$$

under the same boundary conditions. We will seek the solution in the form

$$y_2 = A(x^2 - x) + B(x^3 - x^2) + C(x^4 - x^3). \qquad (14)$$

The coefficients A, B, and C will be determined by Galerkin's method. Finding y_3', y_3'' and substituting them into the left-hand side of equation (13), we obtain

$$L(y_3) = 2Af_1(x) + 2Bf_2(x) + 2Cf_3(x) - x,$$

where

$$f_1(x) = e^{x^3}(4x^2 - 2x + 2) - e^x(x^2 - x),$$
$$f_2(x) = e^{x^3}(6x^3 - 4x^2 + 6x - 2) - e^x(x^3 - x^2),$$
$$f_3(x) = e^{x^3}(8x^4 - 6x^3 + 12x^2 - 6x) - e^x(x^4 - x^3).$$

From the condition of orthogonality of $L(y_3)$ to the functions $\varphi_1(x) = x^2 - x$, $\varphi_2(x) = x^3 - x^2$, and $\varphi_3(x) = x^4 - x^3$ on the interval $[0, 1]$ we obtain the system for determining

the coefficients A, B, and C:

$$A\int_0^1 f_1(x)\,\varphi_1(x)\,dx + B\int_0^1 f_2(x)\,\varphi_1(x)\,dx + C\int_0^1 f_3(x)\,\varphi_1(x)\,dx = \frac{1}{2}\int_0^1 x\varphi_1(x)\,dx,$$

$$A\int_0^1 f_1(x)\,\varphi_2(x)\,dx + B\int_0^1 f_2(x)\,\varphi_2(x)\,dx + C\int_0^1 f_3(x)\,\varphi_2(x)\,dx = \frac{1}{2}\int_0^1 x\varphi_2(x)\,dx, \quad (15)$$

$$A\int_0^1 f_1(x)\,\varphi_3(x)\,dx + B\int_0^1 f_2(x)\,\varphi_3(x)\,dx + C\int_0^1 f_3(x)\,\varphi_3(x)\,dx = \frac{1}{2}\int_0^1 x\varphi_3(x)\,dx.$$

We readily evaluate the integrals on the left-hand sides of system (15):

$$z_1 = \frac{1}{2}\int_0^1 x\varphi_1(x)\,dx = \frac{1}{2}\int_0^1 x(x^2-x)\,dx = \frac{1}{24},$$

$$z_2 = \frac{1}{2}\int_0^1 x\varphi_2(x)\,dx = \frac{1}{2}\int_0^1 x(x^3-x^2)\,dx = \frac{1}{40},$$

$$z_3 = \frac{1}{2}\int_0^1 x\varphi_3(x)\,dx = \frac{1}{2}\int_0^1 x(x^4-x^3)\,dx = \frac{1}{60}.$$

Evaluating the integrals

$$y_{ij} = \int_0^1 f_i(x)\,\varphi_j(x)\,dx, \quad i, j = 1, 2, 3,$$

we obtain: $y_{11} = 0.593292$, $y_{12} = 0.353539$, $y_{13} = 0.240814$, $y_{21} = 0.353539$, $y_{22} = 0.286047$, $y_{23} = 0.225300$, $y_{31} = 0.240814$, $y_{32} = 0.225300$, $y_{33} = 0.104969$.

Substituting the found values of the integrals y_{ij} and z_i into system (15), we reduce it to the form

$$Ay_{11} + By_{12} + Cy_{13} = z_1,$$
$$Ay_{21} + By_{22} + Cy_{23} = z_2,$$
$$Ay_{31} + By_{32} + Cy_{33} = z_3.$$

This system is solved with respect to A, B, and C also on a digit computer. We get: $A = x_1 = -8.678800$, $B = x_2 = 34.396296$, and $C = x_3 = -28.942178$.

Substituting the found coefficients A, B, and C into formula (14), we obtain an approximation for the function minimizing functional (12) or representing the solution of the boundary-value problem for equation (13). It has the form

$$y = x(x - 1)(-8.68 + 34.40x - 28.94x^2).$$

4. Kantorovich's Method. The method of solving variation problems suggested in 1933 by L. Kantorovich occupies an intermediate position between the exact solution of a variation problem and an approximate solution by Ritz' or Galerkin's method. It is also known as the *method of reducing to ordinary differential equations*. It is applicable to functionals of a function of several variables. In addition to a high accuracy, another advantage of this method consists in that instead of the linear combination $\sum_{k=1}^{n} c_k \varphi_k(x, y)$, $c_k = \text{const}$, of the functions $\varphi_k(x, y)$ chosen *a priori*, the solution is sought in the form $\sum_{k=1}^{n} \alpha_k(x) \varphi_k(x, y)$, where $\alpha_k(x)$ are certain functions defined in accordance with the character of the problem solved.

Let, for instance, a variation problem be solved for the functional

$$I(z) = \int\int_G F(x, y, z, z'_x, z'_y)\, dx\, dy$$

provided that on the boundary γ of the domain G the following condition is fulfilled:

$$z(x, y)|_\gamma = 0.$$

For the sake of definiteness, we shall assume that the domain G is bounded by the straight lines $x = a$, $x = b$ and the curves $y = g_1(x)$, $y = g_2(x)$ (Fig. 43). The same as in the Ritz method, a sequence of coordinate functions $\varphi_1(x, y)$, $\varphi_2(x, y)$, ..., ..., $\varphi_n(x, y)$, ... is chosen. The functions vanish on the curves $y = g_1(x)$ and $y = g_2(x)$. We shall seek the solution of the problem in the form

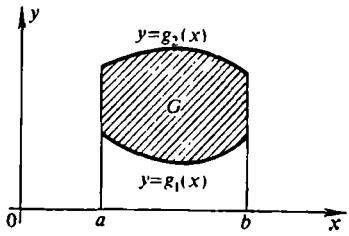


Fig. 43

$$z_n = \sum_{k=1}^{n} \alpha_k(x)\,\varphi_k(x, y),$$

where $\alpha_k(x)$ are unknown differentiable functions satisfying the condition $\alpha_k(a) = \alpha_k(b) = 0$, $k = 1, 2, \ldots, n$. Let us find the partial derivatives

$$\frac{\partial z_n}{\partial x} = \sum_{k=1}^{n} \left(\alpha_k'(x)\,\varphi_k(x, y) + \alpha_k(x)\,\frac{\partial \varphi_k(x, y)}{\partial x}\right),$$

$$\frac{\partial z_n}{\partial y} = \sum_{k=1}^{n} \alpha_k(x)\,\frac{\partial \varphi_k(x, y)}{\partial y}$$

and substitute them into the functional $I(z)$:

$$I(z_n) = \int_a^b dx \int_{g_1(x)}^{g_2(x)} F\left[x, y, \sum_{k=1}^{n} \alpha_k \varphi_k, \sum_{k=1}^{n} \left(\alpha_k' \varphi_k + \alpha_k \frac{\partial \varphi_k}{\partial x}\right), \sum_{k=1}^{n} \alpha_k \frac{\partial \varphi_k}{\partial y}\right] dy.$$

The functions $\varphi_k(x, y)$ are given, and the functions $\alpha_k(x)$ depend only on x, therefore it is possible to perform the integration with respect to y in the inner integral. Then we shall get a function dependent on x, α_k, and α_k', $k = 1$,

2, . . ., n, which will be denoted as

$$\Phi(x, \alpha_1, \alpha_1', \ldots, \alpha_n, \alpha_n') = \int_{g_1(x)}^{g_2(x)} F\left[x, y, \sum_{k=1}^{n} \alpha_k \varphi_k, \sum_{k=1}^{n} \left(\alpha_k' \varphi_k + \alpha_k \frac{\partial \varphi_k}{\partial x}\right), \sum_{k=1}^{n} \alpha_k \frac{\partial \varphi_k}{\partial y}\right] dy.$$

Substituting Φ into the integral $I(z_n)$, we see that

$$I(z_n) = \int_a^b \Phi(x, \alpha_1, \alpha_1', \ldots, \alpha_n, \alpha_n')\, dx = I^*(\alpha_1, \ldots, \alpha_n)$$

becomes a functional of n functions $\alpha_k(x)$ of one variable x. The extremals of this functional satisfy the system of Euler's equations

$$\Phi_{\alpha_k} - \frac{d}{dx} \Phi_{\alpha_k'} = 0, \quad k = 1, 2, \ldots, n, \tag{16}$$

and the boundary conditions $\alpha_k(a) = \alpha_k(b) = 0$.

Thus, in Kantorovich's method solving a variation problem is reduced to solving a system of boundary-value problems for ordinary differential equations.

System (16) can be given in another form which is more suitable for practical applications. For this purpose, it is necessary to find the derivatives $\partial\Phi/\partial\alpha_k$ and $\partial\Phi/\partial\alpha_k'$ of the function

$$\Phi(x, \alpha_1, \alpha_1', \ldots, \alpha_n, \alpha_n') = \int_{g_1(x)}^{g_2(x)} F(x, y, z, z_x', z_y')\, dy,$$

where

$$z = \sum_{k=1}^{n} \alpha_k(x) \varphi_k(x, y), \quad z_x' = \sum_{k=1}^{n} \left(\alpha_k' \varphi_k + \alpha_k \frac{\partial \varphi_k}{\partial x}\right),$$

$$z_y' = \sum_{k=1}^{n} \alpha_k \frac{\partial \varphi_k}{\partial y};$$

we have

$$\frac{\partial \Phi}{\partial \alpha_k} = \int_{g_1(x)}^{g_2(x)} \left(F_z \frac{\partial z}{\partial \alpha_k} + F_{z'_x} \frac{\partial z'_x}{\partial \alpha_k} + F_{z'_y} \frac{\partial z'_y}{\partial \alpha_k} \right) dy$$

$$= \int_{g_1(x)}^{g_2(x)} \left(F_z \varphi_k + F_{z'_x} \frac{\partial \varphi_k}{\partial x} + F_{z'_y} \frac{\partial \varphi_k}{\partial y} \right) dy.$$

The integral $\int_{g_1(x)}^{g_2(x)} F_{z'_y} \frac{\partial \varphi_k}{\partial y} dy$ is transformed by integrating by parts

$$\int_{g_1(x)}^{g_2(x)} F_{z'_y} \frac{\partial \varphi_k}{\partial y} dy = F_{z'_y} \varphi_k (x, y) \Big|_{g_1(x)}^{g_2(x)} - \int_{g_1(x)}^{g_2(x)} \frac{\partial F_{z'_y}}{\partial y} \varphi_k dy.$$

On the curves $y = g_1(x)$ and $y = g_2(x)$ the functions $\varphi_k(x, y)$ vanish, therefore

$$\int_{g_1(x)}^{g_2(x)} F_{z'_y} \frac{\partial \varphi_k}{\partial y} dy = - \int_{g_1(x)}^{g_2(x)} \frac{\partial F_{z'_y}}{\partial y} \varphi_k dy.$$

Thus, the following equalities hold:

$$\frac{\partial \Phi}{\partial \alpha_k} = \int_{g_1(x)}^{g_2(x)} \left[\left(F_z - \frac{\partial F_{z'_y}}{\partial y} \right) \varphi_k + F_{z'_x} \frac{\partial \varphi_k}{\partial x} \right] dy, \quad k = 1, 2, \ldots, n.$$

Let us find the partial derivative

$$\Phi_{\alpha'_k} = \frac{\partial \Phi}{\partial \alpha'_k} = \int_{g_1(x)}^{g_2(x)} F_{z'_x} \varphi_k dy,$$

and for the total derivative we have

$$\frac{d}{dx} \Phi_{\alpha'_k} = \int_{g_1(x)}^{g_2(x)} \left(\varphi_k \frac{\partial}{\partial x} F_{z'_x} + F_{z'_x} \frac{\partial \varphi_k}{\partial x} \right) dy + F_{z'_x} \varphi_k \Big|_{y = g_2(x)} g'_2(x)$$

$$- F_{z'_x} \varphi_k \Big|_{y = g_1(x)} g'_1(x),$$

or, taking into account the boundary conditions, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\Phi_{\alpha_k'} = \int_{g_1(x)}^{g_2(x)} \left(\varphi_k \frac{\partial}{\partial x} F_{z_x'} + F_{z_x'} \frac{\partial \varphi_k}{\partial x}\right) \mathrm{d}y.$$

We then substitute the derivatives $\mathrm{d}\Phi/\mathrm{d}\alpha_k$, $(\mathrm{d}/\mathrm{d}x)\,\Phi_{\alpha_k'}$ into system (16) and reduce it to the form

$$\int_{g_1(x)}^{g_2(x)} \left(F_z - \frac{\partial}{\partial x} F_{z_x'} - \frac{\partial}{\partial y} F_{z_y'}\right) \varphi_k(x,\ y)\,\mathrm{d}y = 0,$$

$$k = 1,\ 2,\ \ldots,\ n. \quad (17)$$

The function in parentheses under the integral sign

$$\Psi(z) = F_z - \frac{\partial}{\partial x} F_{z_x'} - \frac{\partial}{\partial y} F_{z_y'}$$

is the right-hand side of the Ostrogradsky equation for the functional

$$I(z) = \int\int_G F(x,\ y,\ z,\ z_x',\ z_y')\,\mathrm{d}x\,\mathrm{d}y.$$

Thus, for each fixed x conditions (17) may be regarded as the conditions of orthogonality of the functions $\Psi(z)$ to the system of coordinate functions $\varphi_k(x,\ y)$, $k = 1,\ 2,\ \ldots$ $\ldots,\ n$, on the integral $[g_1(x),\ g_2(x)]$.

When finding an extremal by Kantorovich's method, first we have to write the Ostrogradsky equation, and then to solve system (17) which represents a system of ordinary differential equations with respect to the unknown functions $\alpha_k(x)$, $k = 1,\ 2,\ \ldots,\ n$.

In applications, we frequently encounter functionals of the form

$$I(z) = \int\int_G \left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 2zf(x,\ y)\right] \mathrm{d}x\,\mathrm{d}y.$$

System (17) for such a functional has the form

$$\int_{g_1(x)}^{g_2(x)} \left(\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} - f\right) \varphi_k\,\mathrm{d}y = 0, \quad k = 1,\ 2,\ \ldots,\ n, \quad (18)$$

where

$$z = \sum_{k=1}^{n} \alpha_k (x) \varphi_k (x, y).$$

The coordinate system of functions $\varphi_1 (x, y), \ldots$ $\ldots, \varphi_n (x, y), \ldots$ can be constructed in the following way. Let $\psi_1 (y), \ldots, \psi_n (y), \ldots$ be a sequence of coordinate functions defined on [0, 1], and let $g_1 (x)$ and $g_2 (x)$ be the curves determined in Subsection 1. Then $\varphi_n (x, y)$ can be defined as follows:

$$\varphi_n (x, y) = \psi_n \left(\frac{y - g_1 (x)}{g_2 (x) - g_1 (x)} \right), \quad n = 1, 2, \ldots .$$

In particular, for $\varphi_n (x, y)$ we may choose the functions

$$\varphi_n (x, y) = (y - g_1 (x)) (y - g_2 (x)) \cdot y^{n-1}, \quad n = 1, 2, \ldots,$$

or

$$\varphi_n (x, y) = \sin \frac{\pi n (y - g_1 (x))}{g_2 (x) - g_1 (x)}, \quad n = 1, 2, \ldots .$$

Example 4°. Applying Kantorovich's method, find the first approximation of the function minimizing the functional

$$I (z) = \iint_G \left[\left(\frac{\partial z}{\partial x} \right)^2 + \left(\frac{\partial z}{\partial y} \right)^2 + 2z (x + 2) \right] dx\, dy$$

provided that $z (x, y) = 0$ on the boundary γ of the domain G defined by the inequalities $-1 \leqslant x \leqslant 1$ and $-x - 1 \leqslant y \leqslant x + 3$.

According to Kantorovich's method, we seek the first approximation in the form

$$z_1 = (y + x + 1) (y - x - 3) \alpha (x) = [(y - 1)^2 - (x + 2)^2] \alpha (x),$$

where $\alpha (x)$ is an unknown function satisfying the condition $\alpha (-1) = \alpha (1) = 0$. Let us find the partial derivatives

$$\frac{\partial z_1}{\partial x} = [(y - 1)^2 - (x + 2)^2] \alpha' (x) - 2 (x + 2) \alpha (x),$$

$$\frac{\partial^2 z_1}{\partial x^2} = [(y - 1)^2 - (x + 2)^2] \alpha'' (x) - 4 (x + 2) \alpha' (x) - 2\alpha (x),$$

$$\frac{\partial z_1}{\partial y} = 2(y-1)\alpha(x),$$

$$\frac{\partial^2 z_1}{\partial y^2} = 2\alpha(x).$$

The unknown function $\alpha(x)$ is determined from system (18) which in this case is reduced to one equation

$$\int_{-x-1}^{x+3} \{[(y-1)^2-(x+2)^2]\alpha'' - 4(x+2)\alpha' - (x+2)\} \times [(y-1)^2-(x+2)^2]\,dx = 0.$$

Evaluating the integrals

$$\int_{-x-1}^{x+3} dy = 2(x+2), \quad \int_{-x-1}^{x+3} (y-1)^2\,dy = \frac{2}{3}(x+2)^3,$$

$$\int_{-x-1}^{x+3} (y-1)^4\,dy = \frac{2}{5}(x+2)^5,$$

we reduce this equation to the form

$$\frac{16}{15}(x+2)^5\alpha'' + \frac{16}{3}(x+2)^4\alpha' = -\frac{4}{3}(x+2)^4,$$

or, after reduction by (16/15) $(x+2)^4$,

$$(x+2)\alpha'' + 5\alpha' = -\frac{5}{4}. \tag{19}$$

Thus, we have obtained a linear second-order differential equation. We now find the general solution of the corresponding homogeneous equation

$$(x+2)\alpha_0'' + 5\alpha_0' = 0.$$

By making the substitution $\alpha_0' = t$, it is reduced to the first-order equation $(x+2)t' + 5t = 0$, whence we find

$$t = \frac{c_1}{(x+2)^5} \quad \text{or} \quad \alpha_0' = \frac{c_1}{(x+2)^5}.$$

Consequently, $\alpha_0 = -\frac{c_1}{4(x+2)^4} + c_2$. The function $\alpha = -(1/4)(x+2)$ will be the partial solution of equa-

tion (19). Thus the general solution of equation (19) has the form

$$\alpha = c_2 - \frac{c_1}{4(x+2)^4} - \frac{1}{4}(x+2).$$

The constants c_1 and c_2 are found from the boundary conditions $\alpha(-1) = \alpha(1) = 0$. For their determination we have the system

$$c_2 - \frac{c_1}{4} = \frac{1}{4},$$

$$c_2 - \frac{c_1}{4.81} = \frac{3}{4}.$$

Solving this system, we find $c_1 = 81/40$ and $c_2 = 121/160$. Consequently,

$$\alpha(x) = \frac{1}{160}(y+x+1)(y-x-3)\left(41 - 40x - \frac{81}{(x+2)^4}\right),$$

and the first approximation found by Kantorovich's method has the form

$$z_1 = \frac{1}{160}(y+x+1)(y-x-3)\left(41 - 40x - \frac{81}{(x+2)^4}\right).$$

CHAPTER 8

Problems of Computation and Uniform Approximation of Functions

Sec. 8.1. ERRORS DUE TO APPROXIMATE CALCULATIONS

1. Basic Concepts. When constructing a mathematical model of a concrete phenomenon, we have to reduce this phenomenon to mathematical equations, to estimate their parameters, initial data, to carry out analysis, and to choose the method of solution. The mathematical model thus obtained yields only an approximation to the concrete phenomenon and, consequently, describes it with some error. The initial data of a process are usually known not exactly. In most cases, during the process of solving certain equations we have to resort to approximate calculations. Thus, both the construction of a mathematical model and its analysis for the purpose of studying a concrete phenomenon involve mistakes. In order to estimate these mistakes, we introduce the notion of errors which may be either absolute or relative. Let us give their definitions.

Let a be the exact value of a quantity, and let a^* be the value of this quantity obtained as a result of measurements or calculations. Then the *absolute error* of the quantity a^* is defined as the quantity $\Delta(a^*)$ about which it is known that $|a - a^*| \leqslant \Delta(a^*)$, and the *relative error* of the quantity $a^* \neq 0$ is defined as the quantity $\delta(a^*)$ about which it is known that $\left|\frac{a-a^*}{a^*}\right| \leqslant \delta(a^*)$.

According to sources of error, these may be classified under broad headings as follows: irreducible errors, error of the method applied, and computational errors.

The *irreducible error* occurs owing to the inaccuracy of the mathematical model of a process under consideration and

initial data. It is preserved at each step of computations and is transformed in the process of solving a problem, and a knowledge of the magnitude of an irreducible error enables us to estimate the accuracy with which the problem under consideration must be solved.

In order to estimate the effect of the irreducible error on the final result of the process of computation of y, let us represent y in the form of a function

$$y = f(t, p_1, \ldots, p_n, x_1, \ldots, x_m),$$

where t is time, $p_1, \ldots, p_n$ are the parameters of the mathematical model of a given process, and $x_1, \ldots, x_m$ are initial data. Owing to the presence of errors, instead of the quantities $p_1, \ldots, p_n$ and $x_1, \ldots, x_m$, for the parameters of the process and initial data we obtain the values $p_1^*, \ldots, p_n^*$ and $x_1^*, \ldots, x_m^*$, and it is known that

$$|p_i - p_i^*| \leqslant \Delta(p_i^*), \quad i = 1, \ldots, n,$$
$$|x_k - x_k^*| \leqslant \Delta(x_k^*), \quad k = 1, \ldots, m.$$

Therefore for the final result we shall obtain the expression

$$y^* = f(t, p_1^*, \ldots, p_n^*, x_1^*, \ldots, x_m^*),$$

and the absolute error of the result $\Delta(y^*)$ will satisfy the inequality

$$|f(t, p_1^*, \ldots, p_n^*, x_1^*, \ldots, x_m^*) - f(t, p_1, \ldots, p_n, x_1, \ldots, x_m)| \leqslant \Delta(y^*).$$

Remark. If the irreducible errors $\Delta(p_i^*)$ and $\Delta(x_k^*)$ are sufficiently small and the function $f(t, p_1, \ldots, x_m)$ is a differentiable function of the argument $p_1, \ldots, p_n$ and $x_1, \ldots, x_m$, then for the absolute error $\Delta(y^*)$ we may take the quantity

$$\Delta(y^*) = \sum_{i=1}^{n} \left| \frac{\partial f(t, p_1^*, \ldots, x_m^*)}{\partial p_i^*} \right| \Delta(p_i^*) + \sum_{k=1}^{m} \left| \frac{\partial f(t, p_1^*, \ldots, x_m^*)}{\partial x_k^*} \right| \Delta(x_k^*). \quad (1)$$

Although in most cases the function $f(t, p_1, \ldots, x_m)$ cannot be expressed explicitly, expression (1) shows that the error of the result obtained is proportional to the irreducible error.

The *error of the method applied* occurs owing to the fact that the solution of the equations defined by the mathematical model of a given process usually requires an infinite number of operations and turns out to be the limit of a convergent sequence. Confining ourselves to a finite number of operations, we can find only a finite number of terms of this sequence which determine only an approximation to the true solution. The absolute value of the difference between the found approximation and the true solution defines the error of the method applied. Estimation of the error of the method is one of the basic problems of numerical analysis. It will be carried out when considering some concrete methods; here we are not going to dwell on it.

Remark. The error of the method applied can be controlled when solving problems, and, consequently, must be analyzed when synthesizing the algorithm for solving a given problem or during the process of carrying out appropriate computations.

The *computational error* occurs owing to rounding off numbers to a definite number of digits at the input and performing arithmetic operations on the computer. Round-off errors are transformed and accumulated during the process of arithmetic operations. Let us dwell on this question in more detail.

2. Errors Due to Arithmetic Operations. Let us denote by a, b, c the true values of the numbers participating in an arithmetic operation, and by a^*, b^*, c^*, $\Delta(a^*)$, $\Delta(b^*)$, $\Delta(c^*)$ their known approximate values and absolute errors, respectively. Regarding the absolute errors sufficiently small, for the error transformed during arithmetic operations we shall have the following expressions:

$$\begin{aligned}
&c = a + b, \ \Delta(c^*) = \Delta(a^*) + \Delta(b^*),\\
&c = a - b, \ \Delta(c^*) = \Delta(a^*) + \Delta(b^*),\\
&c = ab, \qquad \Delta(c^*) = |a^*|\,\Delta(b^*) + |b^*|\,\Delta(a^*),\\
&c = \frac{a}{b}, \quad b \neq 0, \quad \Delta(c^*) = \frac{\Delta(a^*)}{|b^*|} + \frac{|a^*|\,\Delta(b^*)}{|b^*|^2}.
\end{aligned} \tag{2}$$

In the course of arithmetic operations a partial compensation of errors may occur, therefore expressions (2) yield somewhat excessive estimate for the magnitude of the computational error, but they should be borne in mind especially when performing long chains of computations.

The example below shows how a change in round-off errors affects the computational error.

Example 1°. Compute the number $A = (3 + 2\sqrt{2})^2 (3 - 2\sqrt{2})^2$, taking the numbers 1.4142 and 1.41421 for approximate values of the number $\sqrt{2}$ with round-off errors $2 \cdot 10^{-5}$ and $4 \cdot 10^{-6}$, respectively.

The solution is tabulated below.

Table 1

$\sqrt{2}$	A^*	$\Delta(A^*)$
1.4142	1147.6351	6.3641
1.41421	1152.66682	1.33231

It is seen from the table that in accordance with expressions (2), the computational error is proportional to the round-off errors.

Errors can be compensated for by a rational choice of the order of performing the appropriate operations. Let us show this by computing the same number A, but written in another way.

Example 2°. Compute the number $A = ((3 + 2\sqrt{2})(3 - 2\sqrt{2}))^4$, taking the numbers 1.4142 and 1.41421 for approximate values of the number $\sqrt{2}$.

Now we have the following table:

Table 2

$\sqrt{2}$	A^*	$\Delta(A^*)$
1.4142	1153.3834	0.6157
1.41421	1153.80975	0.09938

Comparing Table 2 with Table 1, we see that a change in the order of computation results in a tenfold decrease of the computational error.

Special attention should be drawn to the cases when we have to perform division by small (by absolute value) numbers or to determine the difference between two large, but little distinguishing numbers. In this case it may happen that the significant digits of the result will be determined by the values of round-off errors or the computational error, which leads to incorrect results. Here is an example.

Example 3°. Solve the system of linear equations

$$\begin{aligned} \sqrt{7}x + \sqrt{2}y &= 1, \\ \sqrt{14}x + \quad 2y &= \sqrt{2}, \end{aligned} \tag{3}$$

rounding off its coefficients to three and to four decimal places.

On rounding off the coefficients, we get the system of equations

$$\begin{aligned} 2.645x + 1.414y &= 1.000, \\ 3.741x + 2.000y &= 1.414, \end{aligned} \tag{4}$$

its solution is: $x = 2.672$ and $y = -4.292$.

If the coefficients of system (3) are computed to four decimal places, then we have the system

$$\begin{aligned} 2.6456x + 1.4142y &= 1.0000, \\ 3.7414x + 2.0000y &= 1.4142, \end{aligned} \tag{5}$$

whose solution is: $x = -0.2246$ and $y = -1.1273$.

Hence it is seen that a change in the accuracy with which the coefficients of system (3) are determined results in a sharp change in the solution. System (3) has the solution $x = c$ and $y = (1/2)(\sqrt{2} - \sqrt{14}c)$, where c is arbitrary, whereas the solutions of systems (4) and (5) are determined by round-off errors.

In Example 3°, we observed the instability of the solution which is caused by the system of equations itself. When carrying out computations on a computer, one has to choose such algorithms for solving problems which elucidate the possibility of sharp fluctuation of the solution caused by a

change in the accuracy of computation and do not allow a considerable distortion of the final result. Note that modern computers, ensuring a high accuracy of computation with a proper choice of a numerical method and the algorithm of its realization, almost always make it possible to reduce the computational error of the final result to a quantity found within the limits of the conditions of a given problem. It is usually supposed that the total error of the final result is made up of the irreducible error, computational error, and the error of the method applied, although, in reality, these errors enter into more complicated relations. If it is known that the method applied for solving a certain problem does not involve an excessive increase in the computational or irreducible error, then we may approximately assume that the total error of the result depends on the error of the method applied.

Sec. 8.2. FUNDAMENTALS OF THE THEORY OF FUNCTION APPROXIMATION

1. Basic Problem. The theory of function approximation was first developed in the works by P. L. Chebyshev. In 1853 he suggested the possibility of approximate representation of a given function by a polynomial of an assigned degree. Later on he invented and developed some original methods for solving the posed problem which also were successively used in some fields of kinematics of mechanisms. "Practice everywhere seeks for the best, for the most advantageous," was the idea which preoccupied him for more than forty years and led to the creation of fundamental direction in the theory of function approximation. Chebyshev's studies were continued by his followers: A. N. Korkin, E. I. Zolotarev, A. A. and V. A. Markovs.

A big role in the development of the theory of function approximation was played by the classical Weierstrass approximation theorem (1885) and also by the fundamental researches of S. N. Bernstein, J. R. Jackson, Ch. J. De la Vallée-Poussin, as well as the Soviet mathematical school. It is worthwhile noting that originally many problems from the theory of function approximation were solved separately (for instance, interpolation and uniform approximation)

and only during recent decades, in connection with a rapid development of functional analysis, some general theories appeared. We shall try to set forth individual questions of the theory of function approximation from the point of view of functional analysis. The general point of the theory of function approximation can be formulated in such a way: How a function, possibly complicated, can be approximately represented with the aid of a simple function, say, with the aid of algebraic or trigonometric polynomials?

It has already been noted that P. L. Chebyshev formulated the problem of representing complicated functions by simpler functions, the latter being usually understood as algebraic or trigonometric polynomials. Later the problems of mathematical physics and many technical applications separated sufficiently broad classes of other functions which were also regarded as "simpler" functions. Further development of functional analysis led to the following formulation of the basic problem of the theory of function approximation.

Let in a Banach space X there be given a set $M \subset X$. It is required to find for any element $x \in X$ an element $y_0 \in M$ such that

$$\inf_{y \in M} \| x - y \|_B = \| x - y_0 \|_B. \qquad (1)$$

The element $y_0 = y_0(x)$ is referred to as the *element of best approximation* of the element $x \in X$.

The statement of this problem gives rise to a number of problems.

(1) Let $x \in X$. Is there in the set $M \subset X$ an element $y_0(x)$ realizing equality (1)? (This is the *problem of the existence of the best approximation element.*)

The validity of setting such a problem is confirmed even by such a simple example. Let M be an open circle $x^2 + y^2 < 1$ in the (x, y)-plane, then for any point $K(x_1, y_1) \notin M$ there is no best element in M.

(2) If in the set M there exists a best approximation element y_0 for every x, then is y unique? (This is the *uniqueness problem.*)

The necessity of such a problem can be illustrated by the following example. Let a plane closed set M have the boundary Γ including a circular arc $\overset{\smile}{AB}$ (Fig. 44). Then for each

point of the line segment OC on the ray O_1C passing through the midpoint D of the chord AB and the centre O of the circle there are two best approximation elements, and for the point O there are infinitely many elements of best approximation.

(3) In the case of the existence and uniqueness of the best approximation element $y_0(x) \in M$, find the algorithm permitting to construct this best approximation or the element $\tilde{y}_0(x)$ which is sufficiently close to the element $y_0(x)$. The measure of proximity is determined by the conditions of a problem under consideration (the *construction problem*).

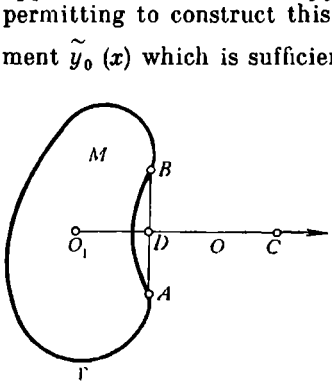


Fig. 44

In what follows, we will concentrate our attention just on these problems, only attributed to concrete spaces X and sets M most frequently used in applications.

2. The Existence Theorem. The main means of approximation in most problems are linear combinations with respect to elements of certain independent systems, therefore in the fundamental problem of the approximation theory the sets of "polynomials" of a definite degree with respect to a linearly independent system of elements of space X are considered as sets $M \in X$.

Problems of such kind are encountered rather frequently, for instance:

(a) An approximate representation of functions with the aid of partial sums of their Taylor series is an approximation with the aid of algebraic polynomials of an assigned degree, the estimate of deviation being uniform on a given interval.

(b) An approximate representation of piecewise smooth periodic functions with the aid of partial sums of their Fourier series is an approximation by means of trigonometric polynomials of a definite degree.

(c) The best mean-square approximation of a square inte-

grable on $[a, b]$ function with the aid of polynomials of an assigned degree with respect to a system of functions orthonormal on $[a, b]$.

Let us first prove the theorem on the existence of the "best" element, that is, the best "polynomial".

Let X be an arbitrary linear normed Banach space (of type B), and let $g_1, g_2, \ldots, g_n$ be n linearly independent elements from X. Then M_n, which is the set of all possible linear combinations, will be:

$$M_n = \{T_n = \sum_{k=1}^{n} \alpha_k g_k\},$$

where α_k, $k = 1, \ldots, n$, are real numbers.

Let x be an arbitrary element from X. The fundamental problem of the theory of approximation can now be formulated as follows. *For a given element $x \in X$ determine the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ so that the quantity*

$$\varphi_x(\alpha_1, \alpha_2, \ldots, \alpha_n) = \| x - \sum_{k=1}^{n} \alpha_k g_k \|_B \qquad (2)$$

receives the least value.

Theorem 1. *For each $x \in X$ there exist numbers $\alpha_1^*, \ldots, \alpha_n^*$ such that the function $\varphi_x(\alpha_1, \alpha_2, \ldots, \alpha_n)$ attains a minimum, that is,*

$$\varphi_x(\alpha_1^*, \ldots, \alpha_n^*) = \inf_{T_n \in M_n} \| x - T_n \|_B.$$

Let us first of all prove that the function $\varphi_x(\alpha_1, \ldots, \alpha_n)$ defined in (2) is a continuous function of its arguments. By virtue of the triangle inequality, we have the relationships

$$| \varphi_x(\alpha_1', \ldots, \alpha_n') - \varphi(\alpha_1, \ldots, \alpha_n) | = | \| x - \sum_{\nu=1}^{n} \alpha_\nu' g_\nu \|$$

$$- \| x - \sum_{\nu=1}^{n} \alpha_\nu g_\nu \| | \leqslant \| \sum_{\nu=1}^{n} (\alpha_\nu' - \alpha_\nu) g_\nu \| \leqslant \sum_{\nu=1}^{n} | \alpha_\nu' - \alpha_\nu | \, \| g_\nu \|$$

$$\leqslant \max_{1 \leqslant \nu \leqslant n} | \alpha_\nu' - \alpha_\nu | \sum_{\nu=1}^{n} \| g_\nu \|$$

meaning that for a sufficiently small $\max\limits_{1\leqslant\nu\leqslant n} |\alpha'_\nu - \alpha_\nu|$ the difference $|\varphi_x(\alpha'_1, \ldots, \alpha'_n) - \varphi_x(\alpha_1, \ldots, \alpha_n)|$ is also small. Let us introduce the function

$$\psi(\alpha_1, \ldots, \alpha_n) = \varphi_\theta(\alpha_1, \ldots, \alpha_n), \qquad \theta = (0, 0, \ldots, 0).$$

The sphere S of elements $\alpha = (\alpha_1, \ldots, \alpha_n)$ such that $\sum\limits_{\nu=1}^{n} |\alpha_\nu|^2 = 1$ is a closed bounded set in the n-dimensional Euclidean space. From the continuity of the functions $\psi(\alpha_1, \ldots, \alpha_n)$ on this sphere, by Weierstrass' theorem, we conclude that the function ψ attains on S a minimal value which we denote by m, that is,

$$m = \min_{(\alpha_1, \ldots, \alpha_n)\in S} \psi(\alpha_1, \ldots, \alpha_n).$$

From the linear independence of the functions $\{g_\nu\}$ it follows that $m > 0$. Using this notation for any point $\beta = (\beta_1, \ldots, \beta_n)$ with the norm $\|\beta\| = \sqrt{\sum\limits_{\nu=1}^{n} |\beta_\nu|^2}$, we may write the relationships

$$\psi(\beta_1, \ldots, \beta_n) = \|\beta_1 g_1 + \ldots + \beta_n g_n\| = \|\beta\| \left\| \frac{\beta_1}{\|\beta\|} g_1 + \ldots + \frac{\beta_n}{\|\beta\|} g_n \right\| \geqslant m\|\beta\|. \quad (3)$$

Suppose now that $\rho_x = \inf \varphi_x(\beta_1, \ldots, \beta_n)$. If $\|\beta\| > (1/m)(\rho_x + 1 + \|x\|) = R$, then from (3) there follows the estimate

$$\varphi_x(\beta_1, \ldots, \beta_n) \geqslant \|\beta_1 g_1 + \ldots + \beta_n g_n\| - \|x\| = \psi(\beta) - \|x\| \geqslant m \cdot \frac{1}{m}(\rho_x + 1 + \|x\|) - \|x\| = \rho_x + 1.$$

Hence, to determine the lower bound of φ_x, we may confine ourselves to the points $\beta = (\beta_1, \ldots, \beta_n)$ lying in a sphere of finite radius R, that is, $\|\beta\| \leqslant R$. But in a bounded closed domain the continuous function $\varphi_x(\beta)$ reaches its minimum, that is, there is a point $\beta^* = (\beta_1^*, \ldots, \beta_n^*)$,

$\| \beta^* \| \leqslant R$, such that $\min \varphi_x(\beta) = \varphi_x(\beta^*)$. In other words, for the element $x \in X$ there exists a polynomial $T_n^* = \sum_{\nu=1}^{n} \beta_\nu^* g_\nu$ such that

$$\| T_n^* - x \| = \inf_{T_n \in M_n} \| T_n - x \|.$$

Sec. 8.3. POLYNOMIALS OF BEST APPROXIMATION IN SPACE $C[a, b]$

1. Properties. Let H_n be a set of polynomials of degree not exceeding n, that is,

$$H_n = \{P_n(x) = \sum_{k=0}^{n} C_k x^k\}. \tag{1}$$

For each function $f(x) \in C[a, b]$ and any polynomial $P(x) \in H_n$ let us consider the deviation $\Delta(P, f) = \| f(x) - P(x) \|_{C[a, b]}$ and choose one of those polynomials $P_n^*(x) \in H_n$ for which

$$\| f(x) - P_n^*(x) \|_C = \inf_{P \in H_n} \Delta(P, f) = \inf_{P \in H_n} \| f(x) - P(x) \|_C.$$

The existence of such polynomials $P_n^*(x)$ is guaranteed by Theorem 1 from the preceding section. The quantity

$$E_n(f) = \| f(x) - P_n^*(x) \|_C = \inf_{P \in H_n} \Delta(P, f)$$

will be called the *best approximation of the function* $f(x)$ by algebraic polynomials of degree no higher than n.

Theorem 1 (Vallée-Poussin's Theorem). *Let for the function* $f(x) \in C[a, b]$ *there exist in* H_n *a polynomial* $\widetilde{P}(x)$ *such that at points* $a \leqslant x_0 < x_1 < \ldots < x_{n+1} \leqslant b$ *the differences* $f(x_k) - \widetilde{P}(x_k)$, $k = 0, 1, \ldots, n+1$, *attain nonzero values* $a_0, a_1, \ldots, a_{n+1}$ *of alternating signs, that is,*

$$(-1)^k [\widetilde{P}(x_k) - f(x_k)] \operatorname{sgn} [\widetilde{P}(x_0) - f(x_0)] = |a_k| > 0. \tag{2}$$

Then for the best approximation $E_n(f)$ *to this function we have the estimate*

$$E_n(f) \geqslant A = \min \{|a_0|, |a_1|, \ldots, |a_n|\}. \tag{3}$$

Let us assume the contrary, i.e. that $E_n(f) < A$. Introducing the notation $\alpha = \operatorname{sgn}[\widetilde{P}(x_0) - f(x_0)]$, we can note that for any $k = 0, 1, \ldots, n+1$ the following relationships are fulfilled:

$$(-1)^k \alpha [\widetilde{P}(x_k) - P_n^*(x_k)] = (-1)^k \alpha [\widetilde{P}(x_k) - f(x_k)] - (-1)^k \alpha [P_n^*(x_k) - f(x_k)] \geqslant A - E_n(f) > 0.$$

Consequently, the difference $\widetilde{P}(x_k) - P_n^*(x_k)$ changes sign on the interval $[a, b]$ at least $n+1$ times, and since this difference is a polynomial of degree no higher than n, this is impossible. Thus, we have arrived at a contradiction.

In order to obtain further properties of the best approximations $E_n(f)$ and polynomials of best approximation $P_n^*(x)$, let us introduce the following definition.

Let $f(x) \in C[a, b]$ and $P(x) \in H_n$. Since the difference $g(x) = P(x) - f(x)$ is continuous on $[a, b]$, there exists at least one point $\widetilde{x} \in [a, b]$ such that $|g(\widetilde{x})| = \Delta(P, f)$. This point $\widetilde{x}$ will be called an e-*point*. According to the sign of $g(\widetilde{x})$, e-points will be distinguished between positive points such that $g(\widetilde{x}) = \Delta(P, f)$, and negative points such that $g(\widetilde{x}) = -\Delta(P, f)$.

Theorem 2 (Chebyshev's Theorem). *For a polynomial $P_n^*(x) \in H_n$ to be a polynomial of best approximation to a continuous function $f(x) \in C[a, b]$, it is necessary and sufficient that there exist on $[a, b]$ at least $m = n + 2$ e-points $x_0 < x_1 < \ldots < x_{n+1}$ which are alternately positive and negative points of the difference $P_n^*(x) - f(x)$.*

Sufficiency. Let the points $x_0 < x_1 < \ldots < x_{n+1}$ be alternately positive and negative points of the difference $P_n^*(x) - f(x)$. Then, setting $\alpha = \operatorname{sgn}[P_n^*(x_0) - f(x_0)]$, we may write

$$(-1)^k \alpha [P_n^*(x_k) - f(x_k)] = \Delta(P_n^*, f), \quad k = 0, 1, \ldots, n+1.$$

Hence, applying Theorem 1, we conclude that $E_n(f) \geqslant \Delta(P_n^*, f)$. But owing to the definition of best approxima-

tion,

$$E_n(f) = \inf_{P \in H_n} \{\Delta(P_n, f)\} \leqslant \Delta(P_n^*, f).$$

From these two inequalities we derive the equality

$$\Delta(P_n^*, f) = E_n(f),$$

that is, $P_n^*(x)$ is the polynomial of best approximation to the function $f(x)$.

Necessity. Suppose the given polynomial $P_n^*(x)$ is the polynomial of best approximation to the function $f(x)$ on the interval $[a, b]$. Since the function $q(x) = P_n^*(x) - f(x)$ is uniformly continuous on $[a, b]$, there is a number $\delta > 0$ such that

$$|q(x') - q(x'')| < \frac{E_n}{2} \text{ if only } |x' - x''| < \delta. \qquad (4)$$

Let us break the interval $[a, b]$ by the division points $a = a_0 < a_1 < \ldots < a_k = b$ into subintervals (or segments) Δ_i, $i = 1, \ldots, k$, the length of each segment being less than δ and choose those segments which contain at least one e-point.

The segments containing at least one positive point will be called positive segments, and those containing at least one negative point will be termed negative segments. Condition (4) implies that one segment cannot contain a positive and a negative point simultaneously.

Let M_1 be the set made up of all positive and negative segments, and let M_2 be the set formed from the remaining segments. We now choose from the set M_1 a segment Δ_{m_0} having the least number and assume, for the sake of definiteness, that it contains a positive point. From the portion of the set M_1 situated to the right of Δ_{m_0} we choose a negative segment Δ_{m_1} having the least number; from the portion of the set M_1 lying to the right of Δ_{m_1} we choose a positive segment Δ_{m_2} with the least number. Since the set M_1 consists of a finite number of segments, continuing this process, we shall separate from this set a finite sequence of segments $\Delta_{m_0}, \Delta_{m_1}, \ldots, \Delta_{m_l}$ which are, alternately, positive and negative segments. Let us assume that the theorem is incorrect, then $l \leqslant n$. Owing to the choice of the quantity δ,

positive and negative segments cannot be adjacent, therefore the segments $\Delta_{m_1-1}, \ldots, \Delta_{m_l-1}$ belong to the set M_2. Let us take on each of these segments Δ_{m_j-1} one interior point $x^{(j)}$, $j = 1, \ldots, l$. The method of constructing the sequence $\Delta_{m_0}, \Delta_{m_1}, \ldots, \Delta_{m_l}$ implies that the portion of the set M_1 to the left of the point $x^{(1)}$ contains only positive segments, and the portion of the set M_1 between the points $x^{(1)}$ and $x^{(2)}$ encloses only negative segments, and so on. Therefore the polynomial of degree $l \leqslant n$

$$R(x) = (-1)^l (x - x^{(1)})(x - x^{(2)}) \ldots (x - x^{(l)}) \qquad (5)$$

will be positive on any positive segment and negative on any negative segment, that is, the signs of the difference $P_n^*(x) - f(x)$ and polynomial $R(x)$ coincide on M_1. Therefore, choosing λ small enough to satisfy the inequality

$$0 < \lambda \mid R(x) \mid < E_n(f), \qquad \forall x \in M_1, \qquad (6)$$

for all $x \in M_1$ we shall have the relationship

$$\mid P_n^*(x) - f(x) - \lambda R(x) \mid = \mid P_n^*(x) - f(x) \mid - \lambda \mid R(x) \mid < E_n(f). \qquad (7)$$

Let us introduce the numbers

$$L = \max_{x \in M_2} \mid P_n^*(x) - f(x) \mid, \quad N = \max_{x \in M_2} \mid R(x) \mid.$$

The set M_2 contains not a single positive or negative segment, therefore $L < E_n(f)$. Consequently, it is possible to choose a positive number λ sufficiently small to fulfill both condition (6) and the inequality $L + \lambda N < E_n(f)$. Then on the set M_2 we shall have the estimate

$$\mid P_n^*(x) - f(x) - \lambda R(x) \mid \leqslant \mid P_n^*(x) - f(x) \mid + \lambda \mid R(x) \mid \leqslant L + \lambda N < E_n(f). \qquad (8)$$

Combining inequalities (7) and (8), we obtain

$$\mid P_n^*(x) - \lambda R(x) - f(x) \mid < E_n(f) \text{ for all } x \in [a, b]. \qquad (9)$$

But for $l \leqslant n$ we have $P_n^*(x) - \lambda R(x) \in H_n$, and inequality (9) contradicts the fact that $P_n^*(x)$ is the polynomial of best approximation. Consequently, $l \geqslant n + 1$.

Corollary. *If $P_n^*(x)$ is the polynomial of best approximation to the function $f(x) \in C[a, b]$, then for $n \geqslant 1$ there exist both positive and negative points.*

In this case $l \geqslant 2$ and the segment Δ_{m_0} contains a positive point, and the segment Δ_{m_1} a negative point.

Theorem 3. *The polynomial $P_n^*(x)$ of best approximation to the function $f(x)$ in H_n is unique.*

Suppose there are two polynomials of best approximation: $P_n^*(x)$ and $Q_n^*(x)$. Then for $x \in [a, b]$ the following relationships are fulfilled:

$$-E_n(f) \leqslant P_n^*(x) - f(x) \leqslant E_n(f),$$
$$-E_n(f) \leqslant Q_n^*(x) - f(x) \leqslant E_n(f),$$

whence we get the inequalities

$$-E_n(f) \leqslant \frac{P_n^*(x) + Q_n^*(x)}{2} - f(x) \leqslant E_n(f).$$

Hence,

$$R_n(x) = \frac{P_n^*(x) + Q_n^*(x)}{2} \in H_n,$$

that is, $R_n(x)$ is a polynomial of best approximation. By Chebyshev's theorem, for $R_n(x)$ there is a system of points $x_1 < x_2 < \ldots < x_{n+2}$ which are alternately positive and negative. Let x_k be a positive point of the difference $R_n - f$, that is,

$$\frac{P_n^*(x_k) - f(x_k)}{2} + \frac{Q_n^*(x_k) - f(x_k)}{2} = E_n(f). \tag{10}$$

It is known that $Q_n^*(x_k) - f(x_k) \leqslant E_n(f)$, therefore formula (10) yields

$$\frac{P_n^*(x_k) - f(x_k)}{2} + \frac{E_n(f)}{2} \geqslant E_n(f),$$

hence it follows that

$$\frac{P_n^*(x_k) - f(x_k)}{2} \geqslant \frac{E_n(f)}{2}. \tag{11}$$

Consequently, at positive points of $(R_n - f)$ we have the equality

$$P_n^*(x_k) - f(x_k) = Q_n^*(x_k) - f(x_k) = E_n(f).$$

Analogous equalities are also true for negative points. Thus, at $n+2$ points the following equalities are valid:

$$P_n^*(x_k) = Q_n^*(x_k),$$

which is possible only for $P_n^*(x) = Q_n^*(x)$.

A system of points $x_0 < x_1 < \ldots < x_n$ satisfying conditions (5) is called *Vallée-Poussin's alternance*, and if, in addition, $|a_k| = E_n(f)$ for all $k = 0, 1, \ldots, n+1$, then it is called *Chebyshev's alternance*.

Theorems analogous to Theorems 1, 2, and 3 can be proved for a trigonometric case as well; then Chebyshev's alternance will consist of no less than $2n+2$ points.

Let us give (without proof) some results illustrating the effect of structural properties of a function on the magnitude of its best approximation. To this end, let us introduce the notion of the modulus of continuity.

The modulus of continuity of the function $f(x)$ on the interval $[a, b]$ is defined as the function

$$\omega(\delta, f) = \max_{\substack{|h|<\delta \\ x,\ x+h\in[a,\ b]}} \|f(x+h) - f(x)\|_{C[a,\ b]}.$$

Theorem 4 (Jackson's Theorem). *If $E_n(f)$ is the best approximation to the function $f(x) \in C[a, b]$ by polynomials from H_n, then the following inequality holds true:*

$$E_n(f) \leqslant 12\omega\left(\frac{b-a}{2n}, f\right).$$

Corollary 1. *If $f(x) \in \mathrm{Lip}_M \alpha$, i.e.*

$$|f(x') - f(x'')| \leqslant M|x' - x''|, \quad 0 < \alpha \leqslant 1,$$

then the following inequality is valid:

$$E_n(f) \leqslant 12\left(\frac{b-a}{2}\right)^{\alpha} \frac{M}{n^{\alpha}}.$$

Corollary 2. *If the function $f(x)$ has a bounded derivative $f'(x)$ and $|f'(x)| \leqslant M$, then the following inequality holds:*

$$E_n(f) \leqslant \frac{6(b-a)M}{n}.$$

Theorem 5 (Jackson's Theorem). *If the function $f(x)$ has at least p continuous derivatives on segment $[a, b]$, then for*

$n > p$ the following estimate is valid:

$$E_n(f) \leqslant \frac{C_p (b-a)^p}{n^p} \omega\left(\frac{b-a}{2(n-p)}, f^{(p)}\right),$$

where $\omega(\delta, f^{(p)})$ is the modulus of continuity of the derivative $f^{(p)}(x)$, and C_p is a constant dependent only on p.

Corollary 1. *If, under the conditions of the theorem, it turns out that*

$$f^{(p)}(x) \in \mathrm{Lip}_M \alpha, \quad 0 < \alpha \leqslant 1,$$

then for $n > p$ the following estimate is fulfilled:

$$E_n(f) \leqslant \frac{C'_p (b-a)^{p+\alpha} M}{n^{p+\alpha}},$$

where the constant C'_p depends only on p and α.

Corollary 2. *If $f(x)$ has a bounded derivative $f^{(p+1)}(x)$ and $|f^{(p+1)}(x)| \leqslant M_{p+1}$, then the following estimate is fulfilled:*

$$E_n(f) \leqslant \frac{C''_p (b-a)^{p+1} M_{p+1}}{n^{p+1}},$$

where the constant C''_p depends only on p.

Theorem 6 (Bernstein's Theorem). *If $f(x)$ is an analytic function on the segment $[a, b]$, then $E_n(f) \leqslant Cq^{(n)}$, where $C > 0$ and $0 < q < 1$ are constant numbers. And if the function $f(x)$ is entire, then*

$$\lim_{n \to \infty} \sqrt[n]{E_n(f)} = 0.$$

From the aforegoing estimates it is seen that if the function $f(x)$ is sufficiently smooth, then the best approximation $E_n(f)$ tends to zero very rapidly. Since the values of algebraic polynomials at any point are well calculated on a computer, it is desirable to use the polynomials of best uniform approximation for evaluating the function always when it is necessary to know the value of the function under consideration at sufficiently large number of points.

2. Construction of Polynomials of Best Approximation. In the case of the existence and uniqueness of a polynomial of best approximation the problem of its construction arises. The exact solution of this problem is possible only in sep-

arate cases. Let us consider one of them: determining the polynomials least deviating from zero (see also Part 1, Sec. 7.2). *Find the polynomial of the nth degree*

$$Q_n(x) = x^n - a_1 x^{n-1} - \ldots - a_{n-1}x - a_n \qquad (12)$$

with the coefficient of x^n *equal to unity for which on a given interval* $[a, b]$ *the quantity* $\max\limits_{a \leqslant x \leqslant b} |Q_n(x)|$ *attains the least possible value.*

Let, for the sake of definiteness, the interval $[a, b]$ coincide with the interval $[-1, 1]$. It is readily seen that the posed problem is equivalent to the problem of finding the polynomial

$$P_{n-1}(x) = a_1 x^{n-1} + \ldots + a_{n-1}x + a_n$$

which is the polynomial of the best approximation to the function x^n on the interval $[-1, 1]$. For determining such a polynomial, note that in this case Chebyshev's alternance consists of $n + 1$ points $x_0, \ldots, x_n \in [-1, 1]$. At each point $x_k \in (-1, 1)$ the function $y = Q_n(x)$ has a local extremum equal to $\pm E_n$, therefore at these points

$$Q_n'(x) = 0. \qquad (13)$$

But $Q_n'(x)$ is a polynomial of degree $n - 1$, hence equation (13) has at most $n - 1$ roots $x_k \in (-1, 1)$. Consequently, the end points of the interval $[-1, 1]$ must also enter into Chebyshev's alternance. Hence it follows that the polynomials $(1 - x^2)\,(Q_n'(x))^2$ and $E_n^2 - Q_n^2(x)$ have the same zeros. Therefore the function $y = Q_n(x)$ satisfies the differential equation

$$(1 - x^2)\,(y')^2 = n^2\,(E_n^2 - y^2). \qquad (14)$$

Rewriting this equation in the form

$$\frac{dy}{\sqrt{E_n^2 - y^2}} = \pm \frac{n\,dx}{\sqrt{1 - x^2}},$$

we find the general solution of equation (14):

$$y(x) = \pm E_n \cos(n \arccos x + C).$$

Setting here $x = 1$, we get

$$|y(1)| = E_n\,|\cos C| = E_n,$$

whence $C = 0$. Also setting $t = \arccos x$, from the identity

$$2\cos nt = (\cos t + i \sin t)^n + (\cos t - i \sin t)^n$$

we find that the expression

$$\cos(n \arccos x) = \frac{1}{2}\left[(x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n\right]$$

is a polynomial of degree n. The coefficient of x^n is readily computed; it is equal to 2^{n-1}. Consequently, the polynomial least deviating from zero is determined by the expression

$$Q_n(x) = \frac{1}{2^{n-1}} \cos(n \arccos x), \quad n = 1, 2, \ldots . \tag{15}$$

Polynomials (15) are called *Chebyshev polynomials of the first kind*. They are frequently applied in approximate calculations.

Let us now give one method of approximate construction of a polynomial of best approximation. Let a continuous function $f(x)$ be defined on the interval $[a, b]$. Suppose there is an approximation to the points of Chebyshev's alternance:

$$a \leqslant x_0^{(k)} < x_1^{(k)} < \ldots < x_n^{(k)} < x_{n+1}^{(k)} \leqslant b. \tag{16}$$

Let us denote by $l_j^{(k)}(x)$ the polynomials of degree n:

$$l_j^{(k)}(x) = \frac{(x - x_1^{(k)}) \ldots (x - x_{j-1}^{(k)})(x - x_{j+1}^{(k)}) \ldots (x - x_{n+1}^{(k)})}{(x_j^{(k)} - x_1^{(k)}) \ldots (x_j^{(k)} - x_{j-1}^{(k)})(x_j^{(k)} - x_{j+1}^{(k)}) \ldots (x_j^{(k)} - x_{n+1}^{(k)})},$$
$$j = 1, 2, \ldots, \quad n+1. \tag{17}$$

Note that if $m = 1, 2, \ldots, n+1$, then

$$l_j^{(k)}(x_m^{(k)}) = \begin{cases} 1, & m = j, \\ 0, & m \neq j. \end{cases}$$

Let us denote by $Q_k(x)$ the polynomial

$$Q_k(x) = \sum_{j=1}^{n+1} [f(x_j^{(k)}) + (-1)^j L_k]\, l_j^{(k)}(x), \tag{18}$$

where L_k is a constant number. If we take the following value of L_k:

$$L_k = \frac{f(x_0^{(k)}) - \sum_{j=1}^{n+1} f(x_j^{(k)}) l_j^{(k)}(x_0^{(k)})}{\sum_{j=1}^{n+1} (-1)^j l_j^{(k)}(x_0^{(k)}) - 1}, \qquad (19)$$

then the polynomial $Q_k(x)$ will satisfy the condition

$$Q_k(x_j^{(k)}) = f(x_j^{(k)}) + (-1)^j L_k, \qquad j = 0, 1, \ldots, n+1,$$

therefore the following equalities are valid:

$$(-1)^j \operatorname{sgn} L_k [Q_k(x_j^{(k)}) - f(x_j^{(k)})] = |L_k|,$$
$$j = 0, 1, \ldots, n+1.$$

Consequently, by Vallée-Poussin's theorem, $E_n(f) \geqslant |L_k|$ and the polynomial $Q_k(x)$ may be chosen for an approximate representation of the polynomial of best approximation.

Let us introduce the notation

$$f_k(x) = \operatorname{sgn} L_k [Q_k(x) - f(x)].$$

Let $x_0^{(k+1)}$ be the point belonging to the segment $[a, x_1^{(k)}]$ at which the function $f_k(x)$ reaches the greatest value, $x_1^{(k+1)}$ be the point of the segment $[x_0^{(k+1)}, x_2^{(k)}]$ at which $f_k(x)$ attains the least value, $x_2^{(k+1)}$ be the point of the segment $[x_1^{(k+1)}, x_3^{(k)}]$ at which $f_k(x)$ reaches its greatest value, and so on. Let $x_{n+1}^{(k+1)}$ be the point of the segment $[x_n^{(k+1)}, b]$ at which $f_k(x)$ reaches the greatest value for odd n, and the least value for even n. The sequence of the points $x_0^{(k+1)}, \ldots, x_{n+1}^{(k+1)}$ thus constructed is taken for a new approximation to the points of Chebyshev's alternance. Indeed, according to the method of choice of the points $x_0^{(k+1)}, x_1^{(k+1)}, \ldots, x_{n+1}^{(k+1)}$ we have the inequality

$$(-1)^j \operatorname{sgn} L_k [Q_k(x_j^{(k+1)}) - f(x_j^{(k+1)})] \geqslant |L_k|,$$

therefore, replacing in (19) the points $x_j^{(k)}$ by the points $x_j^{(k+1)}$, $j = 0, 1, \ldots, n+1$, we find the number L_{k+1} such

that

$$|L_k| \leqslant |L_{k+1}| \leqslant E_n(f).$$

The process is then repeated. The iterative process comes to an end if the quantity $|L_{k+1} - L_k|$ becomes sufficiently small.

As the initial points of approximation to Chebyshev's alternance, it is recommended to take the points

$$x_l = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{n+1-l}{n+1}\pi\right), \quad l = 0, 1, \ldots, n-1, \tag{20}$$

which are the points of Chebyshev's alternance for the polynomial least deviating from zero on the interval $[a, b]$. The described iterative process converges uniformly with a rate defined by the inequality

$$\max_{a \leqslant x \leqslant b} |f(x) - Q_m(x)| \leqslant E_n(f) + cq^m,$$

where $c > 0$, $0 < q < 1$, are certain constants independent of m.

CHAPTER 9

Interpolation and Its Application to Problems of Numerical Differentiation and Integration

Sec. 9.1. INTERPOLATION

1. Interpolation Problem. Let H be a compact set in a normed linear space E, and let

$$f_0(P),\ f_1(P),\ f_2(P),\ \ldots,\ f_n(P),\quad P \in H, \tag{1}$$

be linearly independent functions continuous on H.

The linear combination of these functions

$$F_n(P,\ C) = c_0 f_0(P) + c_1 f_1(P) + \ldots + c_n f_n(P), \tag{2}$$

where c_i, $i = 0, 1, \ldots, n$, are real numbers, will be called the *generalized polynomial*. Let us also assume that a continuous function $f(P)$ is defined on the compact H, and its values $f(P_0)$, $f(P_1)$, $\ldots$, $f(P_n)$ at certain points $P_0, P_1, \ldots, P_n$ are known.

Let us formulate the general problem of interpolation: *Find the generalized polynomial of form* (2) *attaining at the given points* $P_0, P_1, \ldots, P_n$ *of the compact* H *the values equal to* $f(P_0)$, $f(P_1)$, $\ldots$, $f(P_n)$. The points $P_0, P_1, \ldots, P_n$ are called *interpolation points*, and the generalized polynomial $F_n(P, C^*)$ satisfying the conditions

$$F_n(P_k,\ C^*) = f(P_k),\quad k = 0, 1, \ldots, n, \tag{3}$$

is termed the *interpolation polynomial*.

In order for a problem of interpolation to have a solution for any function $f(P)$ continuous on the compact H and for any choice of distinct interpolation points, it is necessary

that the system of linear algebraic equations

$$C_0 f_0(P_k) + C_1 f_1(P_k) + \ldots + C_n f_n(P_k) = f(P_k), \quad k = 0, 1, \ldots, n, \quad (4)$$

equivalent to conditions (3) be solvable. The solvability of system (4) is equivalent to the fact that for any distinct points $P_0, P_1, \ldots, P_n \in H$ the following relationship must be fulfilled:

$$D\begin{pmatrix} f_0, & f_1, & \ldots, & f_n \\ P_0, & P_1, & \ldots, & P_n \end{pmatrix} = \begin{vmatrix} f_0(P_0) & f_1(P_0) & \ldots & f_n(P_0) \\ \ldots & \ldots & \ldots & \ldots \\ f_0(P_n) & f_1(P_n) & \ldots & f_n(P_n) \end{vmatrix} \neq 0.$$

With the determinant equal to zero, the system of equations

$$c_0 f_0(P_k) + c_1 f_1(P_k) + \ldots + c_n f_n(P_k) = 0, \quad k = 0, 1, \ldots, n,$$

has a nontrivial solution $c_0 = b_0, c_1 = b_1, \ldots, c_n = b_n$, and, consequently, the generalized polynomial $\Phi(P) = b_0 f_0(P) + b_1 f_1(P) + \ldots + b_n f_n(P)$ vanishes on the compact H at least at $n + 1$ distinct points $P_0, P_1, \ldots, P_n$. Hence it follows that in order for $D\begin{pmatrix} f_0, f_1, \ldots, f_n \\ P_0, P_1, \ldots, P_n \end{pmatrix} \neq 0$ for any choice of distinct points $P_0, P_1, \ldots, P_n$ of the compact H, it is necessary and sufficient that any generalized polynomial $F_n(P, C)$ different from the identical zero have no more than n zeros on the compact H.

The system of functions (1) defined on the compact H is called the *Chebyshev system of order n* if any generalized polynomial $F_n(P, C)$ having more than n zeros on the compact H is identically equal to zero. From the foregoing there follows the validity of the following theorem.

Theorem 1. *In order for the problem of interpolation for any choice of distinct interpolation points $P_0, P_1, \ldots, P_n$ to have a unique solution, it is necessary and sufficient that system* (1) *be a Chebyshev system.*

Let us now pass to constructing the interpolation polynomial $F_n(P, C^*)$. Let $f_0(P), f_1(P), \ldots, f_n(P)$ be a Chebyshev system on the compact H. We write the system of equa-

tions (4) and equality (2) in the form

$$\begin{aligned} &\sum_{\nu=0}^{n} c_\nu f_\nu(P_k) - f(P_k)\, c_{n+1} = 0, \quad k = 0, 1, \ldots, n, \\ &\sum_{\nu=0}^{n} c_\nu f_\nu(P) - F_n(P, C^*)\, c_{n+1} = 0, \; c_{n+1} = 1. \end{aligned} \tag{5}$$

But this is a system of $n + 2$ homogeneous equations with the unknowns $c_0, c_1, \ldots, c_n, c_{n+1}$. The solvability of system (4) implies the solvability of system (5), and therefore its determinant is equal to zero. Expanding this determinant in the elements of the last column, we obtain the interpolation polynomial

$$F_n(P, C^*) = \sum_{k=0}^{n} f(P_k) \frac{D\begin{pmatrix} f_0, & \ldots, & f_{k-1}, & f_k, & f_{k+1}, & \ldots, & f_n \\ P_0, & \ldots, & P_{k-1}, & P, & P_{k+1}, & \ldots, & P_n \end{pmatrix}}{D\begin{pmatrix} f_0, & f_1, & \ldots, & f_n \\ P_0, & P_1, & \ldots, & P_n \end{pmatrix}} = \sum_{k=0}^{n} f(P_k)\, l_k(P), \tag{6}$$

where

$$l_k(P_j) = \frac{D\begin{pmatrix} f_0, & \ldots, & f_{k-1}, & f_k, & f_{k+1}, & \ldots, & f_n \\ P_0, & \ldots, & P_{k-1}, & P_j, & P_{k+1}, & \ldots, & P_n \end{pmatrix}}{D\begin{pmatrix} f_0, & f_1, & \ldots, & f_n \\ P_0, & P_1, & \ldots, & P_n \end{pmatrix}} = \begin{cases} 0 \text{ for } j \neq k, \\ 1 \text{ for } j = k. \end{cases}$$

It is clear that $l_k(P)$, $k = 0, 1, \ldots, n$, are generalized polynomials on the compact H.

Relationship (6) shows that the interpolation polynomial $F_n(P, C^*)$ represents a linear operator mapping the space $C(H)$ of functions continuous on the compact H onto the finite-dimensional space generated by the generalized polynomials $F_n(P, C)$ defined on the same compact H.

Let now on the compact $H \subset E$ there be given an infinite system of continuous functions $f_0(P), f_1(P), f_2(P), \ldots, f_n(P), \ldots$ such that any of its finite subsystem is a Chebyshev system. Then, given an infinite triangular ma-

trix T of points $P_{kj} \in H$, that is, the matrix

$$T = \begin{pmatrix} P_{00} & & & \\ P_{10} & P_{11} & & \\ \cdots & \cdots & \cdots & \\ P_{n0} & P_{n1} & \ldots & P_{nn} \\ \cdots & \cdots & \cdots & \end{pmatrix},$$

for any function $f(P)$ continuous on H we can construct a sequence of interpolation polynomials $\{F_n(P, C^{(n)})\}_{n=0}^{\infty}$ such that the points of interpolation of the polynomial $F_n(P, C^{(n)})$ will be represented by the points of the nth row of the matrix T, that is,

$$F_n(P_{nj}, C^{(n)}) = f(P_{nj}), \quad j = 0, 1, \ldots, n.$$

Since at the points $P \in H$ different from the points P_{nj}, $j = 0, 1, \ldots, n$, the difference $F_n(P, C^{(n)}) - f(P)$ is, generally speaking, different from zero, there arises the problem of investigating the behaviour of the quantity

$$\varepsilon(F_n, f) = \max_{P \in H} | F_n(P, C^{(n)}) - f(P) | = \| F_n(P, C^{(n)}) - f(P) \|_{C(H)} \quad (7)$$

as $n \to \infty$, and this is the problem of approximation of the function $f(P)$ continuous on H by the sequence of linear interpolation operators $\{F_n(P, C^{(n)})\}$. The question of the possibility of such an approximation can be solved with the aid of the profound theorem of functional analysis which we give without proof.

Theorem 2 (Banach-Steinhaus Theorem). *Let $\{A_n\}$ be a sequence of linear operators mapping a Banach space X into a Banach space Y. In order that for any $x \in X$ the sequence $\{A_n x\}$ converge to the value Ax, where A is a linear operator, it is necessary and sufficient that:*

(a) *the sequence of the norms $\{\| A_n \|\}$ be bounded, i.e. there is a number $M > 0$ such that*

$$\| A_n \|_1 = \sup_{\| x \|_B \leqslant 1} \| A_n x \| \leqslant M \text{ for all } n = 1, 2, \ldots; \quad (8)$$

(b) *for any element z from a set $K \subset X$ whose linear combinations of elements lie everywhere dense in X the following*

relationship hold:

$$\| A_n z - Az \|_B \to 0 \quad as \quad n \to \infty.$$

Condition (8) of boundedness of the norms of the sequence of operators is a sufficiently limiting condition. Thus, it is not fulfilled even for such classical apparatus of approximation in the space of continuous functions as the partial sums of Fourier's series, partial sums of series of orthogonal polynomials given below, Lagrange's interpolation polynomials, and for a number of other methods.

From the Banach-Steinhaus theorem it is possible to readily deduce the Du Bois-Reymond assertion on the existence of a continuous function of period 2π with a nonuniformly convergent or even divergent at separate points Fourier series. A similar assertion can also be obtained for some other approximation methods. Therefore the investigation of the behaviour of the quantity $\varepsilon(F_n, f)$ defined in (7) is reduced to both studying the norm sequence $\{\| F_n \|_1\}$ and finding out the properties of the functions $f(P)$ affecting the rate of convergence of the quantity $\varepsilon(F_n, P)$ to zero. The latter question for concrete interpolation polynomials will be answered in the next subsections.

2. Lagrange's Interpolation Polynomial. Let the function $f(x)$ be uniquely defined and continuous on the interval $[a, b]$, the values $f(x_0), f(x_1), \ldots, f(x_n)$ of this function at some points $a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b$ being known. In this case the system of functions $1, x, x^2, \ldots, x^n$ is taken most often as a Chebyshev system used for constructing the interpolation polynomial. The system of equations (4) then takes the form

$$\sum_{\nu=0}^{n} a_\nu x_k^\nu = f(x_k), \quad k = 0, 1, \ldots, n. \tag{9}$$

The determinant of this system

$$\Delta(x_0, x_1, \ldots, x_n) \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix}$$

is known as a *Vandermonde determinant.* In linear algebra it is proved that

$$\Delta(x_0, x_1, \ldots, x_n) = \prod_{i>j} (x_i - x_j), \tag{10}$$

and since $x_i \neq x_j$ for $i \neq j$, we have $\Delta(x_0, x_1, \ldots, x_n) \neq 0$, whence it follows that the solution of system (9) is existent and unique.

To determine the interpolation polynomial $P_n(x)$, let us take advantage of formula (6). Since in this case

$$l_k(x) = \frac{\Delta(x_0, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n)}{\Delta(x_0, x_1, \ldots, x_n)},$$

using formula (10) and introducing the notation

$$\omega(x) = \prod_{\nu=0}^{n} (x - x_\nu), \tag{11}$$

we may write

$$l_k(x) = \frac{\omega(x)}{\omega'(x_k)(x - x_k)}. \tag{12}$$

Substituting this expression into (6), we obtain *Lagrange's interpolation polynomial*

$$P_n(x) = \sum_{k=0}^{n} f(x_k) \frac{\omega(x)}{\omega'(x_k)(x - x_k)}. \tag{13}$$

Let us now assume that on the interval under consideration the function $f(x)$ has continuous derivatives up to the $(n+1)$th order inclusively. Then it is possible to estimate the error due to replacing the function $f(x)$ by its interpolation polynomial $P_n(x)$. Let us choose the constant R so that the following equality is fulfilled:

$$f(x') - \sum_{k=0}^{n} f(x_k) l_k(x') = R\omega(x'), \tag{14}$$

where x' is a fixed value of x from $[a, b]$ different from the interpolation points x_ν.

The function

$$\Phi(x) = f(x) - \sum_{k=0}^{n} f(x_k) l_k(x) - R\omega(x)$$

vanishes on $[a, b]$ at least $n + 2$ times (at points x', x_0, x_1, . . ., x_n). By virtue of the corollary of Rolle's theorem, there is a point $c \in (a, b)$, $c = c(x')$ such that $\Phi^{(n+1)}(c) = 0$. But differentiating $\Phi(x)$, we get the equality

$$\Phi^{(n+1)}(x) = f^{(n+1)}(x) - R(n+1)!.$$

Setting in this equality $x = c$, we find

$$R = \frac{1}{(n+1)!} f^{(n+1)}(c), \quad c = c(x').$$

Substituting this expression into equality (14), at any point x different from the interpolation points we have the relationship

$$f(x) = \sum_{k=0}^{n} f(x_k) l_k(x) + \frac{f^{(n+1)}(c)}{(n+1)!} \omega(x), \quad c = c(x). \qquad (15)$$

Expression (15) represents Lagrange's interpolation formula with the remainder term

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} \omega(x), \quad c = c(x).$$

Note that formula (15) is obtained for a definite system of interpolation points $x_0, x_1, \ldots, x_n$. If a new point $x_{n+1} \in [a, b]$ different from the previous points is added to this system of points, then both the function $\omega(x)$ and the functions $l_k(x)$ are completely changed, that is, a total recalculation of all the coefficients of the interpolation formula and a new estimate of the remainder term are required. There arises a question of interpolation polynomials free of this shortage. Newton's interpolation polynomials are the desired polynomials.

3. Newton's Interpolation Polynomials. Let us first introduce the following notions. Let the function f be defined and continuous on the interval $[a, b]$ and let u and v be two distinct points of this interval $(u \neq v)$. The *first-order divided difference* for the function $f(x)$ denoted by $f(u, v)$ is defined by

$$f(u, v) = \frac{f(u) - f(v)}{u - v}. \qquad (16)$$

Let $w \in [a, b]$, $w \neq u$ and $w \neq v$. We write the first-order divided difference for the function $f(x)$ using the points v and w. The *second-order divided difference* for the function $f(x)$ denoted by $f(u, v, w)$ is defined by the ratio

$$f(u, v, w) = \frac{f(u, v) - f(v, w)}{u - w}. \tag{17}$$

Analogously, the *n-order divided difference* for the function $f(x)$ with respect to distinct points $x_0, x_1, \ldots, x_{n-1}, x_n$ of the interval $[a, b]$ denoted by $f(x_0, x_1, \ldots, x_{n-1}, x_n)$ is defined as the ratio

$$f(x_0, x_1, \ldots, x_{n-1}, x_n) = \frac{f(x_0, x_1, \ldots, x_{n-1}) - f(x_1, \ldots, x_{n-1}, x_n)}{x_0 - x_n}. \tag{18}$$

From ratios (16)-(18) it is seen that the divided difference $f(x_0, x_1, \ldots, x_n)$ satisfies the relation

$$f(x_0, x_1, \ldots, x_{n-1}, x_n) = f(x_n, x_1, \ldots, x_{n-1}, x_0).$$

Theorem 3. *A divided difference of order n for the polynomial $P(x)$, of degree no higher than n, is a constant quantity.*

Let us write for the polynomial $P(x)$ the first-order divided difference for the points x and x_0 $(x \neq x_0)$:

$$P(x, x_0) = \frac{P(x) - P(x_0)}{x - x_0}.$$

According to Bézout's theorem, the numerator of this ratio is divisible (without a remainder) by the denominator, therefore $P(x, x_0)$ is a polynomial of degree $n - 1$. Analogously the second-order divided difference

$$P(x, x_0, x_1) = \frac{P(x, x_0) - P(x_0, x_1)}{x - x_1}$$

is a polynomial of degree $n - 2$. Continuing this reasoning, we obtain that the divided difference of the nth order is a polynomial of degree zero, that is, a constant quantity.

In order to construct Newton's interpolation polynomial, we find the divided difference of the nth order of Lagrange's interpolation polynomial defined by equality (13). By Theorem 3, this divided difference of order n will be a constant, hence, for any $x \in [a, b]$ we may write the equal-

ity

$$P_n(x_0, x_1, \ldots, x_{n-1}, x) = P_n(x_0, x_1, \ldots, x_{n-1}, x_n). \tag{19}$$

Applying to the divided difference on the left formula (18), we get in succession:

$$P_n(x_0, x_1, \ldots, x_{n-1}, x_n) = -\frac{P_n(x_0, \ldots, x_{n-1})}{x - x_{n-1}} + \frac{P_n(x_0, \ldots, x_{n-2}, x)}{x - x_{n-1}}$$

$$= -\frac{P_n(x_0, \ldots, x_{n-1})}{x - x_{n-1}} + \frac{1}{x - x_{n-1}}\left(-\frac{P_n(x_0, \ldots, x_{n-2})}{x - x_{n-2}} + \frac{P_n(x_0, \ldots, x_{n-3}, x)}{x - x_{n-2}}\right) = -\frac{P_n(x_0, \ldots, x_{n-1})}{x - x_{n-1}} - \frac{P_n(x_0, \ldots, x_{n-2})}{(x - x_{n-1})(x - x_{n-2})}$$

$$- \ldots - \frac{P_n(x_0, x_1)}{(x - x_{n-1}) \ldots (x - x_1)} - \frac{P_n(x_0)}{(x - x_{n-1}) \ldots (x - x_0)} + \frac{P_n(x)}{(x - x_{n-1}) \ldots (x - x_0)}. \tag{20}$$

Conditions (9), i.e. equality $P_n(x_k) = f(x_k)$, imply that

$$P_n(x_0) = f(x_0), \; P_n(x_0, x_1, \ldots, x_m) = f(x_0, x_1, \ldots, x_m), \quad m = 1, \ldots, n.$$

Multiplying (20) by $(x - x_0) \ldots (x - x_{n-1})$ and taking into account (19), we obtain

$$P_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1) + \ldots + f(x_0, x_1, \ldots, x_n)(x - x_0)(x - x_1) \ldots (x - x_{n-1}). \tag{21}$$

The interpolation polynomial $P_n(x)$ written in form (21) is called *Newton's interpolation polynomial* for arbitrary interpolation points.

In practice, we frequently encounter the case of *equally spaced interpolation points*: $x_k = x_0 + kh$, $k = 0, 1, \ldots, n$. The interpolation polynomials obtained in this case are most convenient for computation. For their construction, let us introduce the notion of a finite difference.

The *first-order finite difference* is defined by the expression

$$\Delta f(x_k) = f(x_{k+1}) - f(x_k). \tag{22}$$

The *finite difference of order n* is defined as the first-order finite difference of the finite difference of order $n-1$, that is,

$$\Delta^n f(x_k) = \Delta^{n-1} f(x_{k+1}) - \Delta^{n-1} f(x_k), \quad n = 2, 3, \ldots . \tag{23}$$

Let us establish the relationship between the divided and finite differences:

$$f(x_1, x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{\Delta f(x_0)}{h},$$

$$f(x_2, x_1, x_0) = \frac{f(x_2, x_1) - f(x_1, x_0)}{x_2 - x_0} = \frac{1}{2h}\left(\frac{\Delta f(x_1)}{h} - \frac{\Delta f(x_0)}{h}\right) = \frac{\Delta^2 f(x_0)}{h^2},$$

$$f(x_3, x_2, x_1, x_0) = \frac{f(x_3\ x_2, x_1) - f(x_2, x_1, x_0)}{x_3 - x_0} = \frac{1}{3h}\left(\frac{\Delta^2 f(x_1) - \Delta^2 f(x_0)}{2h^2}\right) = \frac{\Delta^3 f(x_0)}{3!\, h^3}.$$

Using the induction method, one can readily get convinced that for an arbitrary n the following relationship holds true:

$$f(x_n, x_{n-1}, \ldots, x_1, x_0) = \frac{\Delta^n f(x_0)}{n!\, h^n}, \quad n = 1, 2, \ldots . \tag{24}$$

Taking advantage of (24), from (21) we obtain the formula

$$P_n(x) = f(x_0) + \frac{\Delta f(x_0)}{h}(x - x_0) + \frac{\Delta^2 f(x_0)}{2!\, h^2}(x - x_0)(x - x_1) + \ldots + \frac{\Delta^n f(x_0)}{n!\, h^n}(x - x_0)(x - x_1)\ldots(x - x_{n-1}); \tag{25}$$

and setting here $x = x_0 + th$, we find

$$P_n(x_0 + th) = f(x_0) + \sum_{k=1}^{n} \frac{\Delta^k f(x_0)}{k!}\, t(t-1)\ldots(t-k+1). \tag{26}$$

The interpolation polynomials (25) and (26) are called *Newton's interpolation polynomials for equal intervals.*

4. Interpolation with Multiple Points. Approximation by Splines. Up till now, when constructing interpolation polynomials, it was assumed that the interpolation points are distinct. Let us now consider the construction of an interpolation polynomial $H(x)$ of degree $N = \sum_{\nu=0}^{n} \alpha_\nu - 1$, where $\alpha_\nu \geqslant 1$, satisfying the conditions

$$H^{(k_i)}(x_i) = f^{(k_i)}(x_i), \quad i = 0, 1, \ldots, n; \quad k_i = 0, 1, \ldots, \alpha_i - 1, \quad a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b. \quad (27)$$

This problem is known as the *problem of interpolation with multiple points*, and the polynomial $H(x)$ satisfying conditions (27) is called the *Hermite interpolation polynomial*.

The polynomial $H(x)$ is usually sought for in the form

$$H(x) = P_0(x) + (x-x_0)^{\alpha_0} P_1(x) + (x-x_0)^{\alpha_0}(x-x_1)^{\alpha_1} P_2(x) + \ldots + (x-x_0)^{\alpha_0}(x-x_1)^{\alpha_1}\ldots(x-x_{n-1})^{\alpha_{n-1}} P_n(x), \quad (28)$$

where

$$P_k(x) = a_0^{(k)} + a_1^{(k)}(x-x_k) + \ldots + a_{\alpha_k-1}^{(k)}(x-x_k)^{\alpha_k-1}. \quad (29)$$

From (27) and (28) it is easy to note that

$$P_0(x) = f(x_0) + f'(x_0)(x-x_0) + \ldots + \frac{f^{(\alpha_0-1)}(x_0)}{(\alpha_0-1)!}(x-x_0)^{\alpha_0-1},$$

and with the polynomials $P_0(x), \ldots, P_{k-1}(x)$ being known the coefficients of the polynomial $P_k(x)$ are found one by one by successive computation of the derivatives of the expression

$$\frac{H(x) - P_0(x) - (x-x_0)^{\alpha_0} P_1(x) - \ldots - (x-x_0)^{\alpha_0}\ldots(x-x_{k-2})^{\alpha_{k-2}} P_{k-1}(x)}{(x-x_0)^{\alpha_0}\ldots(x-x_{k-1})^{\alpha_{k-1}}} = P_k(x) + (x-x_k)^{\alpha_k} P_{k+1}(x) + \ldots + (x-x_k)^{\alpha_k}\ldots(x-x_{n-1})^{\alpha_{n-1}} P_n(x) \quad (30)$$

taken at the point x_k.

Let us now consider the problem concerning the application of Hermite polynomials. Let the function $y = f(x)$ be defined and continuous on the interval $[a, b]$, and let $a = x_0 < x_1 < \ldots < x_n = b$. We then construct the interpolation polynomials $S_i(x)$, $i = 1, \ldots, n$, satisfying the conditions

$$S_i(x_{i-1}) = y_{i-1}, \quad S_i'(x_{i-1}) = m_{i-1}, \quad S_i''(x_{i-1}) = M_{i-1}, \tag{31}$$

$$S_i(x_i) = y_i, \quad S_i'(x_i) = m_i, \quad S_i''(x_i) = M_i, \quad i = 1, \ldots, n,$$

where $y_i = f(x_i)$, and m_i and M_i are certain constants. Using the above-discussed method of constructing the interpolation polynomial in form (28) and setting $x_i - x_{i-1} = h_i$, we find the formula

$$S_i(x) = y_{i-1} + m_{i-1}(x - x_{i-1}) + \frac{1}{2} M_{i-1}(x - x_{i-1})^2 + (x - x_{i-1})^2 [A_0^{(i)} + A_1^{(i)}(x - x_i) + A_2^{(i)}(x - x_i)^2], \tag{32}$$

where

$$A_0^{(i)} = \frac{y_i - y_{i-1} - m_{i-1}h_i - \frac{1}{2} M_{i-1}h_i^2}{h_i^3},$$

$$A_1^{(i)} = \frac{-3(y_i - y_{i-1}) + (m_i + 2m_{i-1})h_i + \frac{1}{2} M_{i-1}h_i^2}{h_i^4}, \tag{33}$$

$$A_2^{(i)} = \frac{6(y_i - y_{i-1}) - 3(m_i + m_{i-1})h_i + \frac{1}{2}(M_i - M_{i-1})h_i^2}{h_i^5}.$$

The function $S(x)$ defined on $[a, b]$ by the equalities

$$S(x) = S_i(x), \quad x_{i-1} \leqslant x \leqslant x_i, \quad i = 1, 2, \ldots, n, \tag{34}$$

for any choice of the numbers m_i and M_i, $i = 1, \ldots, n$, is continuous and has on $[a, b]$ continuous derivatives of the first and second orders.

Let us choose the numbers m_i and M_i arbitrarily so that $A_1^{(i)} = A_2^{(i)} = 0$, $i = 1, 2, \ldots, n$. Then each of the polynomials $S_i(x)$ will become a polynomial of degree no higher

than three. Such a choice leads to a system of $2n$ linear algebraic equations

$$\begin{gathered}(m_i + 2m_{i-1})\, h_i + \frac{1}{2} M_{i-1} h_i^2 = 3(y_i - y_{i-1}), \\ (m_i + m_{i-1})\, h_i - \frac{1}{6}(M_i - M_{i-1})\, h_i^2 = 2(y_i - y_{i-1}), \\ i = 1, \ldots, n,\end{gathered} \tag{35}$$

in $2(n+1)$ unknowns m_i and M_i, $i = 0, 1, \ldots, n$. Since the direct solution of system (35) for small h_i may involve an increase in the computational error, let us eliminate the variables $m_0, m_1, \ldots, m_n$ from equations (35). To this effect, we write two consecutive pairs of equations from system (35):

$$(m_i + 2m_{i-1}) + \frac{1}{2} h_i M_{i-1} = \frac{3(y_i - y_{i-1})}{h_i},$$

$$(m_i + m_{i-1}) - \frac{1}{6} h_i (M_i - M_{i-1}) = \frac{2(y_i - y_{i-1})}{h_i},$$

$$(m_{i+1} + 2m_i) + \frac{1}{2} h_{i+1} M_i = \frac{3(y_{i+1} - y_i)}{h_{i+1}},$$

$$(m_{i+1} + m_i) - \frac{1}{6} h_{i+1} (M_{i+1} - M_i) = \frac{2(y_{i+1} - y_i)}{h_{i+1}},$$

$$i = 1, 2, \ldots, n-1.$$

We then eliminate from these four equations the three unknowns m_{i-1}, m_i, and m_{i+1}:

$$\frac{h_i}{2} M_{i-1} + (h_{i+1} + h_i) M_i + \frac{h_{i+1}}{2} M_{i+1} = 3\left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}\right).$$

But, by virtue of (17), we have the equality

$$\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} = (h_i + h_{i+1})\, f(x_{i-1}, x_i, x_{i+1}),$$

therefore the preceding equation can be written in the form

$$\frac{1}{2} \frac{h_i}{h_i + h_{i+1}} M_{i-1} + M_i + \frac{1}{2} \frac{h_{i+1}}{h_i + h_{i+1}} M_{i+1} = 3f(x_{i-1}, x_i, x_{i+1}), \tag{36}$$

$$i = 1, 2, \ldots, n-1.$$

These expressions represent a system of $n-1$ linear equations in $n+1$ unknowns $M_0, M_1, \ldots, M_n$. The unknowns M_0 and M_n are usually determined proceeding from the conditions of the problem under consideration. System (36) is a system of linear algebraic equations with a predominant main diagonal, and in the course of its solution the computational error does not increase to a considerable extent.

Let us give one more expression for $S_i(x)$. Suppose that the constants $M_0, M_1, \ldots, M_n$ have already been computed. Since the second derivative of the function $S_i(x)$ is linear on the interval (x_{i-1}, x_i), it can be written in the form

$$S_i''(x) = M_{i-1}\frac{x_i - x}{h_i} + M_i\frac{x - x_{i-1}}{h_i}.$$

Integrating twice both members of this equality and computing the integration constants from the conditions $S_i(x_{i-1}) = y_{i-1}$ and $S_i(x_i) = y_i$, we obtain

$$S_i(x) = M_{i-1}\frac{(x_i - x)^3}{6h_i} + M_i\frac{(x - x_{i-1})^3}{6h_i} + \left(y_{i-1} - \frac{M_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(y_i - \frac{M_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}. \quad (37)$$

The function $S(x)$ constructed by formula (34) is called the *cubic spline*.

When approximating by a cubic spline the function $f(x)$ having on the interval $[a, b]$ continuous derivatives at least up to the fourth order inclusively, the following estimates are valid:

$$\max_{a \leqslant x \leqslant b} |f^{(p)}(x) - S_i^{(p)}(x)| = O(\Delta^{4-p}), \quad p = 0, 1, 2, 3,$$

where $\Delta = \max_i h_i$.

The theory of splines, originated not long ago (in 1946) has been effectively developed recently and widely applied in solving various problems of numerical analysis.

Sec. 9.2. FORMULAS FOR NUMERICAL DIFFERENTIATION AND INTEGRATION. ERROR ESTIMATES

1. Numerical Differentiation. Let a function $f(x)$ differentiable at least $n+2$ times be defined on the interval $[a, b]$, and let its values $f(x_0), f(x_1), \ldots, f(x_n)$ at the points $x_0 < x_1 < \ldots < x_n$ of this interval be known. Formulas of the form

$$\tilde{f}^{(m)}(c) = \sum_{j=0}^{n} A_j^{(m)}(c) f(x_j), \quad m \leqslant n, \tag{1}$$

enabling us to determine an approximate value of $\tilde{f}^{(m)}(c)$ of the m-order derivative of the function $f(x)$ at an arbitrary point $c \in [a, b]$ in terms of the values of the function $f(x)$ at the points $x_0, x_1, \ldots, x_n$ are called the *formulas for numerical differentiation*. For the sake of simplicity, we will consider only the formulas of form (1) when for determining the coefficients $A_j^{(m)}(c)$ we apply Lagrange's interpolation polynomial

$$P_n(x) = \sum_{j=0}^{n} f(x_j) l_j(x).$$

Differentiating this equality m times and setting $\tilde{f}^{(m)}(c) = P_n^{(m)}(c)$, we obtain the formula

$$\tilde{f}^{(m)}(c) = \sum_{j=0}^{n} l_j^{(m)}(c) f(x_j), \quad m \leqslant n. \tag{2}$$

The error $R_m(c) = f^{(m)}(c) - \tilde{f}^{(m)}(c)$ is determined from formula (2). Let $1 \leqslant m \leqslant n$. By Taylor's formula we may write

$$f(x_j) = \sum_{k=0}^{n+1} \frac{f^{(k)}(c)}{k!} (x_j - c)^k + \frac{f^{(n+2)}(s_j)}{(n+2)!} (x_j - c)^{n+2},$$

$$s_j = x_j + \theta_j (c - x_j), \quad 0 < \theta_j < 1,$$

therefore

$$P_n(x) = \sum_{k=0}^{n+1} \frac{f^{(k)}(c)}{k!} \sum_{j=0}^{n} (x_j - c)^k l_j(x) + \sum_{j=0}^{n} \frac{f^{(n+2)}(s_j)}{(n+2)!} (x_j - c)^{n+2} l_j(x). \quad (3)$$

Since formula (15) from the preceding section implies the validity of the equalities

$$(x-c)^k = \sum_{j=0}^{n} (x_j - c)^k l_j(x), \quad k = 0, 1, \ldots, n,$$

$$(x-c)^{n+1} = \sum_{j=0}^{n} (x_j - c)^{n+1} l_j(x) + \omega(x),$$

expression (3) can be written in the form

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k - \frac{f^{(n+1)}(c)}{(n+1)!} \omega(x) + \sum_{j=0}^{n} \frac{f^{(n+2)}(s_j)}{(n+2)!} (x_j - c)^{n+2} l_j(x). \quad (4)$$

Differentiating equality (4) m times ($1 \leqslant m \leqslant n$) and setting $x = c$, we get equality

$$P_n^{(m)}(c) = f^{(m)}(c) - \frac{f^{(n+1)}(c)}{(n+1)!} \omega^{(m)}(c) + \sum_{j=0}^{n} \frac{f^{(n+2)}(s_j)}{(n+2)!} (x_j - c)^{n+2} l_j^{(m)}(c), \quad (5)$$

and since $P_n^{(m)}(c) = \tilde{f}_n^{(m)}(c)$, we obtain

$$R_m(c) = \frac{f^{(n+1)}(c)}{(n+1)!} \omega^{(m)}(c) - \sum_{j=0}^{n} \frac{f^{(n+2)}(s_j)}{(n+2)!} (x_j - c)^{n+2} l_j^{(m)}(c).$$

Hence it follows that for the error $R_m(c)$ the following estimate is true:

$$|R_m(c)| \leqslant \frac{|f^{(n+1)}(c)|}{(n+1)!}|\omega^{(m)}(c)| + \frac{M_{n+2}}{(n+2)!}\sum_{j=0}^{n}|x_j-c|^{n+2}|l_j^{(m)}(c)|, \quad (6)$$

where $M_{n+2} = \max\limits_{a\leqslant x\leqslant b}|f^{(n+2)}(x)|$.

If at some interpolation point x_j the first derivative is computed, then estimate (6) can be improved. Indeed, let us choose the number A so that

$$f'(x_j) - P_n'(x_j) = A\omega'(x_j), \quad \omega(x) = \prod_{k=0}^{n}(x-x_k). \quad (7)$$

Such a choice is always possible, since $\omega'(x_j) \neq 0$. Then the function $\varphi(x) = f(x) - P_n(x) - A\omega(x)$ vanishes on the interval $[a, b]$ at least $n+2$ times (at points $x_0, x_1, \ldots, x_n$, at the point x_j—twice). According to the corollary of Rolle's theorem, there is a point $\xi = \xi(x_j) \in (a, b)$ such that $\varphi^{(n+1)}(\xi) = 0$. Differentiating $f(x)$, we obtain

$$\varphi^{(n+1)}(x) = f^{(n+1)}(x) - A(n+1)!.$$

Setting $x = \xi$ in this equality, we find

$$A = \frac{1}{(n+1)!}f^{(n+1)}(\xi), \quad \xi = \xi(x_j),$$

and from expression (7) we get the equality

$$f'(x_j) - \widetilde{f}'(x_j) = \frac{\omega'(x_j)}{(n+1)!}f^{(n+1)}(\xi). \quad (8)$$

Hence we readily derive the estimate

$$|R_1(x_j)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega'(x_j)|. \quad (9)$$

Example 1°. Using equality (5), construct the formulas for numerical differentiation in the case of equally spaced points $x_k = x_0 + kh$, $k = 0, 1, \ldots, n$, for $n = 2$, $n = 3$,

and $c = x_k$.

For the sake of brevity, we introduce the following notation: $f(x_k) = f_k$, $R_m(x_k) = R_{mk}$, $\max\limits_{a \leqslant x \leqslant b} |f^{(l)}(x)| = M_l$. We now have the following expressions:

(1) if $n = 2$, then

$$f'_0 = \frac{1}{2h}(-3f_0 + 4f_1 - f_2) + R_{10}, \quad f'_1 = \frac{1}{2h}(-f_0 + f_2) + R_{11},$$

$$f'_2 = \frac{1}{2h}(f_0 - 4f_1 + 3f_2) + R_{12},$$

$$f''_i = \frac{1}{h^2}(f_0 - 2f_1 + f_2) + R_{1i}, \quad i = 0, 1, \ldots.$$

For the remainder terms R_{ki} of these formulas the following estimates are valid:

$$|R_{1i}| \leqslant \frac{M_3 h^2}{3}, \quad i = 0, 2, \quad |R_{11}| \leqslant \frac{M_3 h^2}{6},$$

$$|R_{2i}| \leqslant M_3 h + \frac{M_4 h^2}{6}, \quad i = 0, 2, \quad |R_{21}| \leqslant \frac{M_4 h^2}{12};$$

(2) if $n = 3$, then

$$f'_0 = \frac{1}{6h}(-11f_0 + 18f_1 - 9f_2 + 2f_3) + R_{10},$$

$$f'_1 = \frac{1}{6h}(-2f_0 - 3f_1 + 6f_2 - f_3) + R_{11},$$

$$f'_2 = \frac{1}{6h}(f_0 - 6f_1 + 3f_2 + 2f_3) + R_{12},$$

$$f'_3 = \frac{1}{6h}(-2f_0 + 9f_1 - 18f_2 + 11f_3) + R_{13},$$

$$f''_0 = \frac{1}{h^2}(2f_0 - 5f_1 + 4f_2 - f_3) + R_{20},$$

$$f''_i = \frac{1}{h^2}(f_{i-1} - 2f_i + f_{i+1}) + R_{2i}, \quad i = 1, 2,$$

$$f''_3 = \frac{1}{h^2}(-f_0 + 4f_1 + 2f_3 - 5f_2) + R_{23},$$

$$f'''_i = \frac{1}{h^3}(-f_0 + 3f_1 - 3f_2 + f_3) + R_{3i}, \quad i = 0, 1, 2, 3.$$

For the remainder terms R_{ki} of these formulas the following estimates hold true:

$$|R_{1i}| \leqslant \frac{1}{4} M_4 h^3 \quad \text{for} \quad i = 0,\ 3,$$

$$|R_{1i}| \leqslant \frac{1}{12} M_4 h^3 \quad \text{for} \quad i = 1,\ 2;$$

$$|R_{2i}| \leqslant \frac{11}{12} M_4 h^2 + \frac{47}{15} M_5 h^3 \quad \text{for} \quad i = 0,\ 3,$$

$$|R_{2i}| \leqslant \frac{1}{12} M_4 h^2 + \frac{1}{60} M_5 h^3 \quad \text{for} \quad i = 1,\ 2;$$

$$|R_{3i}| \leqslant \frac{3}{2} M_4 h + \frac{57}{20} M_5 h^2 \quad \text{for} \quad i = 0,\ 3,$$

$$|R_{3i}| \leqslant \frac{1}{2} M_4 h + \frac{3}{10} M_5 h^2 \quad \text{for} \quad i = 1,\ 2.$$

Note that the diminution of the step (or tabulation interval) h leads to a decrease in the error R_{mi} along with an increase in the computational error, therefore the reduction of h seems to be reasonable only within certain limits. Let us carry out an appropriate analysis for the frequently used formula for numerical differentiation $\tilde{f}_1' = (1/h^2)(f_0 - 2f_1 + f_2)$, with the error of the method $|R_{21}| \leqslant M_4 h^2/12$. Let in computing the values of $f(x)$ the computational error be equal to δ. Assuming for simplicity that the further computation is performed accurately, we obtain the expression for the total error of this formula:

$$\Delta(\tilde{f}_1'') = \frac{4\delta}{h^2} + \frac{M_4 h^2}{12}.$$

The minimum value of $\Delta(\tilde{f}_1'')$ is reached for $h_0 = 2\sqrt[4]{3\delta/M_4}$, and $\min \Delta(\tilde{f}_1'') = \Delta_0 = 2\sqrt{M_4 \delta/3}$.

2. Numerical Integration. Let the function $f(x)$ be defined and continuous on $[a, b]$, and let

$$I(f) = \int_a^b f(x)\,dx. \tag{10}$$

In technical applications, one has frequently to resort to an approximate evaluation of integral (10). This problem is

solved in a most suitable way with the aid of the formulas

$$I(f) \approx S_n(a, b; f) = \sum_{k=0}^{n} A_k^{(n)} f(x_k), \qquad (11)$$

where the constants $A_k^{(n)}$ and the points x_k, $k = 0, 1, \ldots, n$, are independent of the choice of the function $f(x)$.

Formulas of form (11) allowing to determine an approximate value of integral (10) in terms of the value of the function $f(x)$ at fixed points $x_0, x_1, \ldots, x_n$ of the interval $[a, b]$ are called the *quadrature formulas*, the numbers $A_k^{(n)}$, the *weight coefficients*, and the points $x_0, \ldots, x_n$, the *distinct points of the quadrature formula*. The error of the method can be determined by estimating the expression

$$R_n(f) = |I(f) - S_n(a, b; f)|. \qquad (12)$$

The same as in the case of numerical differentiation, we shall confine ourselves to the quadrature formulas (11) obtained with the aid of Lagrange's interpolation polynomials. Let the interval $[a, b]$ be finite, and let

$$P_n(x) = \sum_{k=0}^{n} f(x_k) l_k(x)$$

be Lagrange's interpolation polynomial constructed for the function $f(x)$ using the points $a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b$. Setting approximately

$$I(f) \approx \int_a^b P_n(x)\,dx = \sum_{k=0}^{n} f(x_k) \int_a^b l_k(x)\,dx,$$

we obtain the quadrature formula

$$S_n(a, b; f) = \sum_{k=0}^{n} A_k^{(n)} f(x_k), \qquad (13)$$

where

$$A_k^{(n)} = \int_a^b l_k(x)\,dx, \quad k = 0, 1, \ldots, n. \qquad (14)$$

The quadrature formulas of form (13) in which the weight coefficients $A_i^{(n)}$ are determined by formulas (14) are called

the *interpolation quadrature formulas.* If $f(x) = Q_n(x)$ is a polynomial of degree not exceeding n, then $f(x) = \sum_{k=0}^{n} f(x_k)\, l_k(x)$, and, consequently, $R_n(f) = 0$, that is, the interpolation quadrature formulas (13) yield the exact result at least for all polynomials $Q_n(x)$ of degree no higher than n.

Most frequently, interpolation quadrature formulas are constructed for equally spaced points. The following are the simplest quadrature formulas:

(a) the *rectangular formula*

$$S_0(a, b; f) = (b-a) f\left(\frac{a+b}{2}\right) \tag{15}$$

which is obtained from formula (13) for $n = 0$, $x_0 = (a + b)/2$, $P_0(x) = f(x_0)$. Note that if $f(x) = cx + d$, then obviously, we have the equalities

$$\int_a^b f(x)\, dx = (b-a)\left(c\frac{a+b}{2} + d\right) = S_0(a, b; f),$$

that is, formula (15) yields the exact result for any polynomial of the first degree;

(b) the *trapezoidal formula*

$$S_1(a, b; f) = \frac{b-a}{2}[f(a) + f(b)] \tag{16}$$

which is obtained from (13) for $n = 1$, $x_0 = a$, $x_1 = b$, $P_1(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}$. Note that formula (16) also yields the exact result for any first-degree polynomial;

(c) *Simpson's formula*

$$S_2(a, b; f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{17}$$

which corresponds to the choice $n = 2$, $x_0 = a$, $x_1 = (a + b)/2$, $x_2 = b$ and

$$P_2(x) = \frac{2\left(x - \frac{a+b}{2}\right)(x-b)}{(b-a)^2} f(a) - \frac{4(x-a)(x-b)}{(b-a)^2} f\left(\frac{a+b}{2}\right) + \frac{2(x-a)\left(x - \frac{a+b}{2}\right)}{(b-a)^2} f(b).$$

The interpolation character of Simpson's formula implies that it gives the exact result for any polynomial of the second degree. But it is easily seen that if $f(x) = x^3$, then

$$\int_a^b f(x)\,dx = \frac{b^4 - a^4}{4} = \frac{b-a}{6}\left[a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3\right] = S_2(a, b; f),$$

and, consequently, formula (17) yields the exact result for any polynomial of the third degree.

To improve the accuracy of computation of the definite integral (10) with the aid of a quadrature formula of form (11), the interval $[a, b]$ is divided into m equal parts by the points $\bar{x}_j = a + \frac{b-a}{m} j$, $j = 0, 1, \ldots, m$, and the quadrature formula (13) is applied to each subinterval $(\bar{x}_j, \bar{x}_{j+1})$. The quadrature formula thus constructed

$$I_m(a, b; f) = \sum_{j=0}^{m-1} S_n(\bar{x}_j, \bar{x}_{j+1}; f) = \sum_{j=0}^{m-1} \sum_{k=0}^{n} A_k^{(n)} f(x_{jk}), \quad (18)$$

where $x_{jk} = \bar{x}_j + \frac{\bar{x}_{j+1} - \bar{x}_j}{n} k$, is called the *complicated quadrature formula*. This is how Simpson's complicated quadrature formula looks like:

$$\int_a^b f(x)\,dx = \frac{b-a}{6m}\left[f(\tilde{x}_0) + 4f(\tilde{x}_1) + 2f(\tilde{x}_2) + \ldots + 4f(\tilde{x}_{2m-1}) + f(\tilde{x}_{2m})\right], \quad (19)$$

where $\tilde{x}_i = a + \frac{b-a}{2m} i$, $i = 0, 1, \ldots, 2m$.

3. Error Estimate of a Quadrature Formula. Error estimate of interpolation quadrature formulas can be obtained proceeding from the estimate of the remainder term of Lagrange's interpolation polynomial, but far better results are obtained by the method suggested by the Soviet mathematician S. Nikolsky in his book *Quadrature Formulas* (Moscow, 1976). Let us dwell on the estimate obtained by this method.

We shall say that the function $f(x)$ belongs to the class $W^{(r)}(M_r, a, b)$ if all of its derivatives up to order $r-1$ are continuous on the interval $[a, b]$, the rth derivative is piecewise continuous, and

$$|f^{(r)}(x)| \leqslant M_r, \quad a \leqslant x \leqslant b. \tag{20}$$

Let us denote by R_{nr} the number

$$R_{nr} = \sup_{f(x) \in W^{(r)}(M_r, a, b)} \left| S_n(a, b; f) - \int_a^b f(x)\,dx \right|,$$

and by $F_{nr}(t)$ the function

$$F_{nr}(t) = \frac{(1-t)^r}{r} - \sum_{k=0}^{n} \frac{A_k^{(n)}}{b-a} K_r\left(\frac{x_k - a}{b-a} - t\right), \tag{21}$$

where

$$K_r(u) = \begin{cases} u^{r-1} & \text{for } u \geqslant 0, \\ 0 & \text{for } u < 0. \end{cases}$$

S. Nikolsky showed that if the quadrature formula (11) yields the exact result at least for all polynomials of the $(r-1)$th degree, then

$$R_{nr} = (b-a)^{r+1} M_r c_{nr}, \tag{22}$$

where

$$c_{nr} = \frac{1}{(r-1)!} \int_0^1 |F_{nr}(t)|\,dt. \tag{23}$$

Example 2°. Find the error of the quadrature formula (17) on the class $W^{(4)}(M_4, a, b)$.

According to formulas (17) and (21), we get the equalities

$$F_{24}(t)=\frac{(1-t)^4}{4}-\frac{2}{3}K_4\left(\frac{1}{2}-t\right)-\frac{1}{6}K_4(1-t)$$

$$=\begin{cases}\dfrac{t^3}{4}\left(t-\dfrac{2}{3}\right) & \text{for} \quad 0\leqslant t\leqslant\dfrac{1}{2},\\[2ex] \dfrac{(1-t)^3}{4}\left(\dfrac{1}{3}-t\right) & \text{for} \quad \dfrac{1}{2}\leqslant t<1,\end{cases}$$

and, consequently,

$$c_{24}=\frac{1}{6}\int_0^1 |F_{24}(t)|\,dt$$

$$=\frac{1}{24}\left[\int_0^{1/2}\left(\frac{2}{3}-t\right)t^3\,dt+\int_{1/2}^1\left(t-\frac{1}{3}\right)(1-t)^3\,dt\right]-\frac{1}{2880},$$

therefore, by (22), we have

$$R_{24}=\frac{M_4(b-a)^5}{2880}. \tag{24}$$

Let us apply Nikolsky's method to estimate the error of the complicated quadrature formula (18). To this end, taking advantage of formulas (18) and (21), we obtain the inequalities

$$\left|\int_a^b f(x)\,dx-L_m(a,\ b;\ f)\right|$$

$$\leqslant\sum_{j=0}^{m-1}\sup_{f\in W^{(r)}}\left|\int_{\bar{x}_j}^{\bar{x}_{j+1}} f(x)\,dx-S_n(\bar{x}_j,\ \bar{x}_{j+1};\ f)\right|$$

$$\leqslant\sum_{j=0}^{m-1}\frac{(\bar{x}_{j+1}-\bar{x}_j)^{r+1}M_r}{(r-1)!}\int_0^1|F_{nr}^{(j)}(t)|\,dt, \tag{25}$$

where

$$F_{nr}^{(j)}(t)=\frac{(1-t)^r}{r}-\sum_{k=0}^{n}\frac{A_k^{(n)}}{\bar{x}_{j+1}-\bar{x}_j}K_r\left(\frac{x_{jk}-\bar{x}_j}{\bar{x}_{j+1}-\bar{x}_j}-t\right).$$

Noting that

$$\bar{x}_{j+1} - \bar{x}_j = h = \frac{b-a}{m}, \quad F_{nr}^{(j)}(t) = F_{nr}^{(0)}(t) = \frac{(1-t)^r}{r}$$

$$- \frac{1}{h} \sum_{k=0}^{n} A_k^{(n)} K_r \left(\frac{(x_{0k} - a)}{h} - t \right), \quad j = 0, 1, \ldots, m-1,$$

from inequality (25) we derive the relationship

$$\sup_{f \in W^{(r)}(M_r, a, b)} \left| L_m(a, b; f) - \int_a^b f(x)\, dx \right| \leqslant \frac{h^r (b-a) M_r}{(r-1)!} \int_0^1 |F_{nr}^{(0)}(t)|\, dt. \quad (26)$$

If the function $f(x)$ has continuous derivatives up to the $(r+1)$th order inclusively on the interval $[a, b]$, and the quadrature formula (18) is exact for all the polynomials of degree $r-1$, but does not yield the exact result for rth-degree polynomials, then the following inequality holds true:

$$\left| \int_a^b f(x)\, dx - L_m(a, b; f) - \frac{h^r \varkappa}{(r-1)!} \int_a^b f^{(r)}(t)\, dt \right| \leqslant h^{r+1} (b-a) M_{r+1} c_{nr}, \quad (27)$$

where

$$\varkappa = \int_0^1 F_{nr}^{(0)}(t)\, dt, \quad M_{r+1} = \max_{a \leqslant x \leqslant b} |f^{(r+1)}(x)|,$$

and the constant c_{nr} is defined by equality (23).

The expression

$$h^r G_r = h^r \frac{\varkappa}{(r-1)!} \int_a^b f^{(r)}(t)\, dt \quad (28)$$

is called the *principal error term*.

G. Runge showed that the principal term of the error can be determined in the process of numerical integration.

For this purpose, an approximate value of the integral is computed by the quadrature formulas $L_{m_1}(a, b; f)$ and $L_{m_2}(a, b; f)$ $(m_1 < m_2)$. Inequality (27) implies the equalities

$$\int_a^b f(x)\,dx - L_{m_1}(a, b; f) = h_1^r G_r + \rho_1,$$

$$\int_a^b f(x)\,dx - L_{m_2}(a, b; f) = h_2^r G_r + \rho_2,$$

where

$$h_i = \frac{b-a}{m_i}, \quad |\rho_i| \leqslant h_i^{r+1}(b-a) M_{r+1} c_{nr}, \quad i = 1, 2. \tag{29}$$

Subtracting the second equality from the first, we get

$$G_r(h_1^r - h_2^r) = L_{m_2}(a, b; f) - L_{m_1}(a, b; f) + \rho_2 - \rho_1,$$

therefore

$$h_2^r G_r = \frac{L_{m_2}(a, b; f) - L_{m_1}(a, b; f)}{\left(\frac{h_1}{h_2}\right)^r - 1} + \frac{\rho_2 - \rho_1}{\left(\frac{h_1}{h_2}\right)^r - 1}. \tag{30}$$

Let us estimate the second term of the right-hand side of expression (30). From expression (29) we have the inequalities

$$\left|\frac{\rho_2 - \rho_1}{\left(\frac{h_1}{h_2}\right)^{r-1} - 1}\right| \leqslant h_2^{r+1}(b-a) M_{r+1} c_{nr} \frac{\left(\frac{h_1}{h_2}\right)^{r+1} + 1}{\left(\frac{h_1}{h_2}\right)^r - 1} = O(h_2^{r+1}).$$

Inequality (27) can be transformed to the form

$$\left|\int_a^b f(x)\,dx - L_{m_2}(a, b; f)\right| \leqslant |T| + O(h_2^{r+1}), \tag{31}$$

where

$$T = \frac{L_{m_2}(a, b; f) - L_{m_1}(a, b; f)}{\left(\frac{h_1}{h_2}\right)^r - 1}.$$

We usually take $h_1/h_2 = 2$; the computation process is finished when $|T| < \varepsilon$, where ε is the specified accuracy.

Sec. 9.3. OPTIMIZATION METHODS. CUBATURE FORMULAS

1. Optimization of Quadrature Formulas. Entering upon solving a concrete problem of numerical integration, one has to make up his mind to choose the quadrature formula which should be used to solve the given problem. To choose the most suitable method is a difficult thing, since each method possesses positive and negative sides.

The question of minimizing the number of points of a quadrature formula for ensuring a specified accuracy leads to the following optimization problem: *Given n, construct the quadrature formula of form* (11) (*see the preceding section*)

$$\int_a^b f(x)\,dx \cong S_n(a, b; f) = \sum_{j=0}^{n} A_j^{(n)} f(x_j)$$

exact for polynomials of highest degree. Such quadratures are called *Gaussian formulas.* We shall establish the Gaussian formula in the supposition that the interval $[a, b]$ coincides with the interval $[-1, 1]$. Thus, the desired quadrature formula is defined by the expression

$$\int_{-1}^{1} f(x)\,dx \cong S_n(f) = \sum_{j=0}^{n} A_j^{(n)} f(x_j). \qquad (1)$$

Let us denote by $\omega(x)$ the polynomial of degree $n+1$

$$\omega(x) = (x - x_0)(x - x_1) \ldots (x - x_n).$$

Theorem 1. *In order for the quadrature formula* (1) *to be exact for all the polynomials $P(x)$ of degree less than and equal to m, $m \geqslant n+1$, it is necessary and sufficient that it be an interpolation formula, and the polynomial $\omega(x)$ be orthogonal to all the polynomials $Q(x)$ of degree no higher than $m - n - 1$.*

Necessity. Let $R(x)$ be an arbitrary polynomial of the nth degree. Then, by Lagrange's formula (15) (see Sec. 9.1), we have the equality

$$R(x) - \sum_{j=0}^{n} R(x_j) \frac{\omega(x)}{\omega'(x_j)(x - x_j)},$$

and since, by the conditions of the theorem, when applied to $R(x)$, formula (1) yields the exact result ($n < m$), we have

$$\int_{-1}^{1} R(x)\,dx = \sum_{j=0}^{n} R(x_j) \int_{-1}^{1} \frac{\omega(x)\,dx}{\omega'(x_j)(x-x_j)} = \sum_{j=0}^{n} A_j^{(n)} R(x_j),$$

and, consequently,

$$A_j^{(n)} = \int_{-1}^{1} \frac{\omega(x)\,dx}{\omega'(x_j)(x-x_j)},$$

that is, (1) is an interpolation formula.

Let us now show that $\omega(x)$ is orthogonal to any polynomial $Q(x)$ of degree no higher than $m-n-1$. The product $Q(x)\omega(x)$ is a polynomial of degree no higher than m, and, consequently, when applied to this polynomial, formula (1) yields the exact result, and this means that

$$\int_{-1}^{1} Q(x)\,\omega(x)\,dx = \sum_{j=0}^{n} A_j^{(n)} Q(x_j)\,\omega(x_j) = 0;$$

the orthogonality of $\omega(x)$ and $Q(x)$ has been proved.

Sufficiency. Let $P(x)$ be an arbitrary polynomial of degree no higher than m. Then it can be represented in the form

$$P(x) = Q(x)\,\omega(x) + R(x), \tag{2}$$

where $Q(x)$ and $R(x)$ are polynomials of degree no higher than $m-n-1$ and n, respectively. Taking into consideration the orthogonality of the polynomials $Q(x)$ and $\omega(x)$, we may write

$$\int_{-1}^{1} P(x)\,dx = \int_{-1}^{1} R(x)\,dx. \tag{3}$$

But since (1) is an interpolation formula, and $R(x)$ is a polynomial of degree no higher than n, the following equality holds:

$$\int_{-1}^{1} R(x)\,dx = \sum_{j=0}^{n} A_j^{(n)} R(x_j),$$

It follows from (2) that $R(x_j) = P(x_j)$, $j = 0, 1, \ldots, n$, therefore substituting the value of the integral of $R(x)$ into (3), we get the equality

$$\int_{-1}^{1} P(x)\,dx = \sum_{j=0}^{n} A_j^{(n)} P(x_j),$$

and this means that formula (1) is exact for all the polynomials of degree no higher than m.

Corollary. *The highest degree of the polynomials for which the Gaussian formula* (1) *yields the exact result is equal to* $2n + 1$.

By the conditions of the theorem, any polynomial $Q(x)$ of degree no higher than $m - n - 1$ must be orthogonal to the polynomial $\omega(x)$ of degree $n + 1$. But this is possible only if the condition $m - n - 1 \leqslant n$ is fulfilled, and, consequently, $m \leqslant 2n + 1$.

From the condition of orthogonality of the polynomial $\omega(x)$ to all the polynomials $Q(x)$ of degree no higher than n on the interval $[-1, 1]$ it follows that $\omega(x)$ is a Legendre polynomial (see Part 1, Sec. 7.2). Let us write $\omega(x)$ in the form

$$\omega(x) = \frac{(n+1)!}{(2n+2)!} \frac{d^{n+1}}{dx^{n+1}} (x^2 - 1)^{n+1} \tag{4}$$

and note that all zeros x_j, $j = 0, 1, \ldots, n$, of the Legendre polynomials are distinct and are situated on the interval $(-1, 1)$. It is possible to show that the coefficients $A_j^{(n)}$ of the Gaussian formula (1) are determined by the expressions

$$A_j^{(n)} = \frac{2^{2n+3}[(n+1)!]^4}{[(2n+2)!]^2 (1 - x_j^2)[\omega'(x_j)]^2}, \quad j = 0, 1, \ldots, n, \tag{5}$$

and for the error of the Gaussian formula

$$R_n(f) = \int_{-1}^{1} f(x)\,dx - \sum_{j=0}^{n} A_j^{(n)} f(x_j)$$

the following estimate is valid:

$$|R_n(f)| \leqslant \frac{2^{2n+3}[(n+1)!]^4}{(2n+1)[(2n+2)!]^3} \max_{-1 \leqslant x \leqslant 1} |f^{(2n+2)}(x)|. \quad (6)$$

If the Stirling formula (see Part 1, Sec. 4.2) is used, then estimate (6) will turn into the estimate

$$|R_n(f)| \leqslant \sqrt{\pi}\left(\frac{e}{4}\right)^{2n+2} \times \frac{1+O(1/n)}{(2n+1)(n+1)^{2n+1}\sqrt{n+1}} \max_{-1 \leqslant x \leqslant 1} |f^{(2n+2)}(x)|. \quad (7)$$

Example 1°. Construct the Gaussian quadrature formula for $n = 2$. By formulas (1) and (5), we obtain the equality

$$\int_{-1}^{1} f(x)\,dx = \frac{5}{9} f\left(-\frac{\sqrt{15}}{5}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\frac{\sqrt{15}}{5}\right) + R_2(f), \quad (8)$$

where

$$|R_2(f)| \leqslant \frac{1}{4050} \max_{-1 \leqslant x \leqslant 1} |f^{(6)}(x)|.$$

2. Cubature Formulas. Let the function $f(x_1, \ldots, x_m)$ be defined and differentiable sufficient number of times in a domain G of the m-dimensional Euclidean space E_m. The *cubature formula* is defined as the sum

$$S_N(G, f) = \sum_{k=1}^{N} A_k^{(N)} f(x_1^{(k)}, \ldots, x_m^{(k)}), (x_1^{(k)}, \ldots, x_m^{(k)}) \in \bar{G}, \quad (9)$$

yielding an approximate value of the integral

$$I(f) = \int \ldots \int_{(G)} f(x_1, \ldots, x_m)\,dx_1 \ldots dx_m \cong S_N(G, f).$$

In what follows, for the sake of simplicity, we shall confine ourselves to the case $m = 3$ which already possesses all the peculiarities of numerical computation of many-dimensional integrals. In this case the cubature formula (9) takes the form

$$\int\int\int_{(G)} f(x, y, z)\,dx\,dy\,dz \cong \sum_{k=1}^{N} A_k^{(N)} f(x_k, y_k, z_k). \quad (10)$$

If the domain G is simple with respect to the y- and z-axes, then the triple integral can be reduced to an iterated integral:

$$\iiint\limits_{(G)} f(x, y, z)\,dx\,dy\,dz = \int\limits_a^b dx \int\limits_{y_1(x)}^{y_2(x)} dy \int\limits_{z_1(x,y)}^{z_2(x,y)} f(x, y, z)\,dz. \quad (11)$$

In this case for the numerical integration over the domain G we can make use of the quadrature formulas of form (11) from the preceding section. Let us give the scheme of constructing the cubature formula (10). We first choose a quadrature formula

$$S_n(a, b; f) = \sum_{k=0}^{n} A_k f(x_k), \quad a \leqslant x_0 < \ldots < x_n \leqslant b$$

and determine the numbers $y_{k1} = y_1(x_k)$ and $y_{k2} = y_2(x_k)$. For each pair of such numbers we choose the numbers n_k and construct the formulas

$$S_{n_k}(y_{k1}, y_{k2}; f) = \sum_{j=0}^{n_k} A_{kj} f(y_{kj}), \quad y_{k1} \leqslant y_{kj} \leqslant y_{k2}.$$

Then the numbers $z_{1kj} = z_1(x_k, y_{kj})$ and $z_{2kj} = z_2(x_k, y_{kj})$ are determined. For each pair of such numbers we choose the number n_{kj} and the quadrature formula

$$S_{n_{kj}}(z_{1kj}, z_{2kj}; f) = \sum_{l=0}^{n_{kj}} A_{kjl} f(z_{kj_l}), \quad z_{1kj} \leqslant z_{kj_l} \leqslant z_{2kj}.$$

In this case the cubature formula (10) will have the form

$$S(G, f) = \sum_{k=0}^{n} \sum_{j=0}^{n_k} \sum_{l=0}^{n_{kj}} A_k A_{kj} A_{kjl} f(x_k, y_{kj}, z_{kj_l}). \quad (12)$$

The process of constructing the cubature formula (12) is well algorithmized and can be carried out on a computer without its being constructed beforehand. The error of this formula can be estimated by analogy with the one-dimen-

sional case. Note also that if all the coefficients A_k, A_{kj}, A_{kjl} are positive and the quadrature formulas S_n, S_{n_k}, $S_{n_{kj}}$ yield the exact result, at least for all the constants, then the following formula holds true:

$$\sum_{k=0}^{n} \sum_{j=0}^{n_k} \sum_{l=0}^{n_{kj}} A_k A_{kj} A_{kjl} = V(G),$$

where $V(G)$ is the volume of the domain G. Therefore the application of the cubature formula (12) does not involve an increase in the computational error.

The simplest form is taken by formula (12) when the domain G is a parallelepiped, that is, for $a_1 \leqslant x \leqslant a_2$, $b_1 \leqslant y \leqslant b_2$, $c_1 \leqslant z \leqslant c_2$. Setting $x_k = a_1 + kh_1$, $y_j = b_1 + jh_2$, $z_l = c_1 + lh_3$ and choosing the quadrature formulas

$$S_{n_1}(a_1, a_2; f) = \sum_{k=0}^{n_1} A_k f(x_k), \qquad n_1 = \frac{a_2 - a_1}{h_1},$$

$$S_{n_2}(b_1, b_2; f) = \sum_{j=0}^{n_2} B_j f(y_j), \qquad n_2 = \frac{b_2 - b_1}{h_2},$$

$$S_{n_3}(c_1, c_2; f) = \sum_{l=0}^{n_3} C_l f(z_l), \qquad n_3 = \frac{c_2 - c_1}{h_3},$$

rom expression (12) we have the formula

$$M(G, f) = \sum_{k=0}^{n_1} \sum_{j=0}^{n_2} \sum_{l=0}^{n_3} A_k B_j C_l f(x_k, y_j, z_l). \tag{13}$$

As an example, we give Simpson's formula for a cube with side h

$$M(G, f) = \frac{h^3}{216}(\sigma_1 + 4\sigma_2 + 16\sigma_3 + 64\sigma_4), \tag{14}$$

where σ_1 is the sum of the values of $f(x, y, z)$ at the vertices of the cube, σ_2 the sum of the values of $f(x, y, z)$ at the midpoints of the edges of the cube, σ_3 the sum of the values of $f(x, y, z)$ at the centres of its faces, and σ_4 is the value of $f(x, y, z)$ at the centre of the cube.

In the case of an arbitrary domain G, it is sometimes enclosed in a parallelepiped, and formula (13) is applied for computing the integral of the function

$$F(x, y, z) = \begin{cases} f(x, y, z) & \text{for} \quad (x, y, z) \in \overline{G}, \\ 0 & \text{for} \quad (x, y, z) \notin \overline{G}. \end{cases}$$

But one should be highly circumspect to use this method, since on the boundary of the domain G the function $F(x, y, z)$ may turn out to be discontinuous, that is, the error of the method may turn out to be too large.

3. The Monte Carlo Method. The basic shortage of the foregoing method of constructing cubature formulas consists in a large number of points at which the value of the function $f(x, y, z)$ must be computed. When using formula (13) for the case $n_1 = n_2 = n_3 = n = 10$, this function must be evaluated at 1331 points. With an increase in n the number of points increases considerably, and since the error of the method is determined by the number n, but not n^3, to ensure the specified accuracy, the number n should be chosen sufficiently large. One of the methods in which the error is independent of the dimension of the integral is the *Monte Carlo method.*

Suppose there are N random, pairwise independent points $B_1, \ldots, B_N$ of the m-dimensional Euclidean space equally distributed in the domain G with distribution density $p(x_1, \ldots, x_m)$. Then the following equality is valid:

$$\int_{(G)} \ldots \int p(x_1, \ldots, x_m)\, dx_1 \ldots dx_m = 1.$$

To evaluate the integral

$$I(f) = \int_{(G)} \ldots \int f(x_1, \ldots, x_m)\, dx_1 \ldots dx_m \qquad (15)$$

we may take advantage of the following technique: let us introduce the random quantity $y_j(f)$ by setting

$$y_j(f) = \frac{f(B_j)}{p(B_j)}.$$

The points B_j are pairwise independent, therefore the random quantities $y_j(f)$ will also be pairwise independent, the below relationships being true for the expectation and variance of $y_j(f)$:

$$M(y_j(f)) = \int\limits_{(G)} \dots \int \frac{f(x_1, \dots, x_m)}{p(x_1, \dots, x_m)} p(x_1, \dots, x_m)\, dx_1 \dots dx_m$$

$$= \int\limits_{(G)} \dots \int f(x_1, \dots, x_m)\, dx_1 \dots dx_m = I(f),$$

$$D(y_j(f)) = M(y_j^2(f)) - [M(y_j(f))]^2 = E,$$

where

$$E = I\left(\frac{f^2(x_1, \dots, x_m)}{p^2(x_1, \dots, x_m)}\right) - [I(f)]^2.$$

Suppose the random quantity $U_N(f)$ is defined by the equality

$$U_N(f) = \frac{1}{N} \sum_{j=1}^{N} y_j(f).$$

Taking into account the pairwise independence of the quantities $y_j(f)$, we find the expectation

$$M(U_N(f)) = \frac{1}{N} \sum_{j=1}^{N} M(y_j(f)) = I(f)$$

and variance

$$D(U_N(f)) = \frac{1}{N^2} \sum_{j=1}^{N} D(y_j(f)) = \frac{1}{N} E.$$

Taking advantage of the Chebyshev inequality, we obtain that for any η, $0 < \eta < 1$, the probability of the fulfilment of the inequality $|I(f) - U_N(f)| \leqslant \sqrt{E/\eta N}$ will exceed $1 - \eta$. In this estimate the variance E may be represented by the sample variance

$$E \cong E_s = \frac{1}{N-1} \sum_{j=1}^{N} [y_j(f) - U_N(f)]^2.$$

Thus, computing $I(f)$ by the formula

$$I(f) \cong U_N(f) = \frac{1}{N} \sum_{j=1}^{N} y_j(f)$$

with probability more than, or equal to, $1 - \eta$, we have the estimate

$$|I(f) - U_N(f)| \leqslant \sqrt{E_s/\eta N}.$$

The main difficulty encountered in application of the Monte Carlo method consists in sampling sufficiently long sequences of random points B_j. It is common practice to use pseudorandom numbers for this purpose, but the statistical properties of such sequences must be checked. When computing many-dimensional integrals, the magnitude of the relative error should also be analyzed in order not to allow its excessive increase.

CHAPTER 10

Numerical Methods of Solving Algebraic and Differential Equations

Sec. 10.1. SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS

1. General. Consider a system of n linear equations in n independent variables

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= a_{1,\,n+1} \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= a_{2,\,n+1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= a_{n,\,n+1} \end{aligned} \tag{1}$$

which can be written in matrix form as

$$A\mathbf{x} = \mathbf{b},$$

where $A = (a_{ij})$ is a matrix of dimension $n \times n$, $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbf{b} = (a_{1,\,n+1}, \ldots, a_{n,\,n+1})^T$. Here H^T denotes the matrix obtained by transposing the matrix H. We shall also assume that the matrix A is nonsingular, that is, $\det A \neq 0$.

For solving system (1) exact and iterative methods are used. *Exact methods* are such methods which yield a solution after a finite number of arithmetic operations. If the coefficients and the right-hand sides of system (1) are known exactly and all computations are performed without roundoffs, then the exact solution is obtained. *Iterative methods* are such methods which yield the solution of a problem as the limit of a sequence of approximations computed by a uniform process. When applying iterative methods, of importance is the rapidity of convergence of the constructed approximations. Note that when solving a system of equations on a computer even by exact methods, the solution may be obtained with an error. This happens owing to both the error in putting the coefficients of system (1) in the

computer and the necessity of accomplishing round-offs during the process of solving the system of equations on the computer.

This chapter is dedicated mainly to those methods which are most frequently used in solving large systems of linear equations.

Theoretically, the solution of system (1) is given by the formula

$$\mathbf{x} = A^{-1}\mathbf{b}, \tag{2}$$

where A^{-1} is the *inverse* of A. The matrix A^{-1} is said to be *stable* if small changes in the elements of the matrix A^{-1} correspond to small changes in the elements of the matrix A, otherwise it is *unstable*. From expression (2) it is seen that if the matrix A^{-1} is stable, then small errors of constant terms and coefficients of system (1) cause small changes in the solution $\mathbf{x}$ of this system. But if the matrix A^{-1} is unstable, then small errors of the right-hand sides and coefficients of system (1) can considerably distort its solution.

The matrix A is said to be *poorly conditioned* if the reciprocal matrix A^{-1} is unstable. When solving systems of equations with poorly conditioned matrices, one should be extremely careful, since the error of solution may turn out to be inadmissibly high. Note that although there exist the estimates of matrix conditionality (conditionality numbers), their determination requires much more computational work than the solution of system (1). In order, simultaneously with determining the solution of system (1), to estimate approximately the error of this solution, we proceed as follows: choose a vector $\mathbf{y}^{(0)}$, say $\mathbf{y}^{(0)} = (1, \ldots, 1)^T$, and compute the vector $\mathbf{b}^{(0)} = A\mathbf{y}^{(0)}$. Then at one and the same time, applying one and the same method, we solve the system of equations $A\mathbf{x} = \mathbf{b}$ and $A\mathbf{y} = \mathbf{b}^{(0)}$. The amount of computations here increases, but to an insignificant extent. If it turns out that the ratio $\| \mathbf{y} - \mathbf{y}^{(0)} \| / \| \mathbf{y} \|$ is sufficiently small, then we regard that in solving system (1) the error did not increase considerably. Of course, such reasoning is not rigorous, but the conclusions drawn are sufficiently reliable.

2. Gauss' Elimination Method. The simplest and widely spread method of solving system (1) is *Gauss' method of suc-*

cessive elimination of unknowns. This method has many various computational schemes. Let us consider the unique division scheme.

Suppose that system (1) is such that $a_{11} \neq 0$. Dividing all the coefficients and the constant term of the first equation of system (1) by a_{11}, we get the equation

$$x_1 + a_{12}^{(1)} x_2 + \ldots + a_{1n}^{(1)} x_n = a_{1,n+1}^{(1)}, \quad a_{1j}^{(1)} = \frac{a_{1j}}{a_{11}}. \qquad (3)$$

Then we subtract equation (3) multiplied by the appropriate coefficient a_{i1} from each of the remaining equations of system (1). As a result, we obtain the system

$$\begin{aligned} &a_{i2}^{(1)} x_2 + a_{i3}^{(1)} x_3 + \ldots + a_{in}^{(1)} x_n = a_{i,n+1}^{(1)}, \\ &a_{ij}^{(1)} = a_{ij} - a_{ij}^{(1)} a_{11}, \quad i = 2, 3, \ldots, n. \end{aligned} \qquad (4)$$

Assuming that $a_{22}^{(1)} \neq 0$, we perform an analogous operation. Proceeding in the same way, we obtain the system of equations of the triangular form

$$\begin{aligned} x_1 + a_{12}^{(1)} x_2 + \ldots + a_{1,n-1}^{(1)} x_{n-1} + a_{1n}^{(1)} x_n &= a_{1,n+1}^{(1)}, \\ x_2 + \ldots + a_{2,n-1}^{(2)} x_{n-1} + a_{2n}^{(2)} x_n &= a_{2,n+1}^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ x_{n-1} + a_{n-1,n}^{(n-1)} x_n &= a_{n-1,n+1}^{(n-1)}, \\ x_n &= a_{n,n+1}^{(n)}, \end{aligned} \qquad (5)$$

whose solution is the solution of system (1). Substituting x_n determined by the last equation of system (5) into the last but one equation of the same system, we determine x_{n-1}; knowing x_n and x_{n-1}, we find x_{n-2} from the next equation. This process is repeated until all the components of the solution of system (5) are found.

Thus, to solve system (2) by using the adopted scheme, we first construct an auxiliary triangular system and then solve it. The collection of operations that brings the original system to the triangular form is ordinarily called the *forward procedure of the Gauss' method*, and the set of operations in solving the triangular system, the *backward procedure of this method*. Note also that, simultaneously, we can find

$$\det A = a_{11} a_{22}^{(1)} \ldots a_{nn}^{(n-1)}.$$

The realization of Gauss' elimination method requires the performance of $(1/3)(n^3 + 3n^2 - n)$ operations of multiplication and division.

Note that if at the kth step of the computational process the element $a_{kk}^{(k-1)}$, by which we have to divide all the coefficients of the first equation of the remaining system, equals zero or is too small, then this equation must not be used for eliminating x_k from other equations of the system because of a sharp increase in the computational error. Therefore we have either to interchange the equations of system (1) or to alter the numbers of the sought-for unknowns.

3. Banded Matrices. Sweep Method. If a matrix A is such that its elements a_{ij} satisfy the condition

$$a_{ij} = 0 \text{ for } |i - j| > m, \quad m < n, \tag{6}$$

then A is called the *banded matrix*.

Banded matrices are frequently encountered in solving applied problems. For instance, when determining a cubic spline (see Sec. 9.1) and also when solving boundary-value problems for ordinary partial differential equations, we come to the solution of system (1), when matrix A satisfies condition (6). For all such problems it is characteristic that system (1) should have a large number of equations, but the number m be far less than n. This considerably reduces the total number of arithmetic and logical operations required for solving system (1). It is especially convenient to solve systems of linear equations of form (1) with banded matrices for small values of m.

Consider the case $m = 1$. Then the system of equations (1) can be represented in the form

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 &= d_1,\\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= d_2,\\
a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= d_3,\\
&\cdots\cdots\cdots\cdots\\
a_{n-1,\,n-2}x_{n-2} + a_{n-1,\,n-1}x_{n-1} + a_{n-1,\,n}x_n &= d_{n-1},\\
a_{n,\,n-1}x_{n-1} + a_{nn}x_n &= d_n.
\end{aligned} \tag{7}$$

Let us solve this system by applying Gauss' scheme of unique division. Dividing the first equation by $p_1 = a_{11}$, we

reduce it to the form

$$x_1 - q_1 x_2 = u_1,$$

where $q_1 = -a_{12}/p_1$, $u_1 = d_1/p_1$. Multiplying this equation by a_{21} and subtracting the relationship thus obtained from the second equation of system (7), we get the equation

$$p_2 x_2 + a_{23} x_3 = d_2^{(1)},$$

where $p_2 = a_{22} + a_{21} q_1$, $d_2^{(1)} = d_2 - a_{21} u_1$, of the same form as the first equation of system (7). Thus, the first step of Gauss' method yields a system of the same form as system (7), but containing $n - 1$ equations. As a result of fulfilment of the $(k - 1)$th step of Gauss' method, we arrive at the system

$$\begin{aligned}
p_k x_k + a_{k,\,k+1} x_{k+1} &= d_k^{(1)},\\
a_{k+1,\,k} x_k + a_{k+1,\,k+1} x_{k+1} + a_{k+1,\,k+2} x_{k+2} &= d_{k+1},\\
\cdots\cdots\cdots

which enables us to successively determine $x_{n-1}, \ldots, x_1$.

The method of solving system (7) thus obtained is called the *sweep method* (or the *method of successive substitution*). The process of computing the quantities q_k and u_k from scheme (9) is generally called the *forward sweep procedure*, and the process of solving system (10) is called the *backward sweep procedure*. The sweep method for solving system (7) of order n requires $9n$ arithmetic operations.

4. Iterative Method. The *method of simple iteration* is the most commonly used method in solving the systems of linear equations. Let the system of equations (1) have a unique solution. As is shown in Sec. 3.4, the solution $\mathbf{x}^*$ of system (1) is a fixed point of the operator

$$U\mathbf{x} \stackrel{\text{def}}{=} (I - DA)\,\mathbf{x} + D\mathbf{b} = C\mathbf{x} + \mathbf{b}^*, \quad C = (C_{kj}), \quad (11)$$

where I is a unit matrix, and D is a nonsingular matrix, i.e.

$$\mathbf{x}^* = U\mathbf{x}^* = C\mathbf{x}^* + \mathbf{b}^*. \quad (12)$$

The **iterative process**

$$\mathbf{x}^{(n+1)} = C\mathbf{x}^{(n)} + \mathbf{b}^*, \quad n = 0, 1, \ldots, \quad (13)$$

started with an arbitrary element $\mathbf{x}^{(0)}$ will be convergent if one of the following conditions is fulfilled:

$$\alpha_1 = \max_j \Big(\sum_{j=1}^{n} |C_{kj}|\Big) < 1 \quad \text{(for the space } E_n^*) \quad (14)$$

or

$$\alpha_2 = \Big(\sum_{k=1}^{n}\sum_{j=1}^{n} (C_{kj})^2\Big)^{1/2} < 1 \quad \text{(for the space } E_n). \quad (15)$$

And from the principle of contracted mappings there follows the estimate of approximation

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\|_\nu \leqslant \frac{\alpha_\nu^m}{1 - \alpha_\nu} \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_\nu, \quad \nu = 1, 2. \quad (16)$$

Note that in a concrete process, instead of system (13), we solve the system

$$\tilde{\mathbf{x}} = (C + \Delta C)\,\tilde{\mathbf{x}} + \mathbf{b}^* + \Delta\mathbf{b}^*,$$

where C and $\mathbf{b}^*$ are exact values of initial matrices, and ΔC and $\Delta \mathbf{b}^*$ are round-off errors. Besides, each step of the iterative process is accompanied by rounding off the approximation $\tilde{\mathbf{x}}^{(m)}$ by the quantity $\Delta \mathbf{x}^{(m)}$ ($\| \Delta \mathbf{x}^{(m)} \|_v < \Delta$). If $\| C + \Delta C \|_v \leqslant \alpha_v < 1$ and $\tilde{\mathbf{x}}_1$ is the limit of the constructed sequence of vectors $\tilde{\mathbf{x}}^{(m)}$ (with such round-offs taken into account), then it is not difficult to derive the estimate

$$\| \mathbf{x}^* - \tilde{\mathbf{x}}_1 \| \leqslant \frac{1}{1-\alpha_v} (\| \Delta C \|_v + \| \Delta \mathbf{b}^* \|_v \| \mathbf{x}^* \|_v + \Delta) \quad (17)$$

determining the accuracy with which system (13) can be solved using the method of simple iteration.

Remark. A modification of the method of simple iteration is *Seidel's method.* It consists in that the approximation found for the component x_i is used for determining the next component $x_{i+1}^{(m+1)}$. The appropriate computations are carried out by the formula

$$x_{i+1}^{(m+1)} = \sum_{j=1}^{i} c_{i+1,\,j} x_j^{(m+1)} + \sum_{j=i+1}^{n} c_{i+1,\,j} x_j^{(m)} + b_{i+1}, \quad i = 0, \ \ldots, \ n-1. \quad (18)$$

If $\| C \|_v \leqslant \alpha_v < 1$, then Seidel's method converges somewhat quicker than the method of simple iteration. In the general case, the regions of convergence of Seidel's method and the method of simple iteration are overlapped partially, that is, it is possible to indicate matrices C such that Seidel's method converges, while the method of simple iteration diverges, and vice versa, Seidel's method diverges and the method of simple iteration converges. We are not going to dwell on the proof of these assertions.

5. Eigenvalues and Eigenvectors of Symmetric Matrices. The algebraic problem of eigenvalues consists in determining those values of λ (eigenvalues) for which the system of n homogeneous linear equations in n unknowns

$$A\mathbf{x} = \lambda \mathbf{x}, \quad A = (a_{ij}) \quad (19)$$

has a nontrivial solution $\mathbf{x}$ (eigenvector of the matrix). The problem of finding the eigenvalues and eigenvectors of a matrix has to be resorted to when carrying out the numerical solution of many applied problems, for instance, problems leading to solving systems of linear differential equations with constant coefficients, when solving Fredholm's homogeneous integral equations of the second kind, when investigating linear transformations, and so forth.

In order for system (19) to have a nontrivial solution, it is necessary and sufficient that the equality $\det(A - \lambda I) = 0$ be fulfilled. Expanding the determinant in powers of λ, we obtain the equation

$$\lambda^n + p_1\lambda^{n-1} + \ldots + p_{n-1}\lambda + p_n = 0. \qquad (20)$$

This equation is called the *characteristic equation of the matrix* A, and its roots the *eigenvalues of the matrix* A. Note that for large values of n the coefficients p_i of equation (20) are highly sensitive to small changes in the elements a_{ij} of the matrix A, therefore the eigenvalues of the matrix A are determined directly from equation (20) only when n is small. Confining ourselves to symmetric matrices, let us consider one of the methods enabling us to determine the eigenvalues and eigenvectors of the matrix A without solving the characteristic equation.

Suppose that the matrix A is symmetric ($A^T = A$). In this case A is a self-adjoint operator mapping the n-dimensional Euclidean space E_n into E_n. Therefore for the operator A all the results of Chapter 5 hold true. We will regard that the matrix A has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, and let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ be their corresponding eigenvectors. Then all the eigenvalues of the matrix A are real, and the eigenvectors are orthogonal (see Subsections 3 and 4 of Sec. 5.2). We shall assume that $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$ and that the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ are normed, that is, $(\mathbf{u}_i, \mathbf{u}_i) = 1$. By definition, we have the equalities

$$A\mathbf{u}_k = \lambda_k\mathbf{u}_k, \quad k = 1, \ldots, n. \qquad (21)$$

Let $\mathbf{u}$ be an arbitrary vector. Then it can be uniquely represented in the form $\mathbf{u} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \ldots + a_n\mathbf{u}_n$, where $a_1, a_2, \ldots, a_n$ are, generally speaking, unknown constants. Taking into consideration equality (21), after

p-fold application of the operator A, we get the relationships

$$A^p\mathbf{u} = A^{p-1}\left(\sum_{k=1}^{n} a_k A\mathbf{u}_k\right) = A^{p-1}\left(\sum_{k=1}^{n} a_k\lambda_k\mathbf{u}_k\right)$$

$$= A^{p-2}\left(\sum_{k=1}^{n} a_k\lambda_k^2\mathbf{u}_k\right) = \ldots = \sum_{k=1}^{n} a_k\lambda_k^p\mathbf{u}_k,$$

or

$$A^p\mathbf{u} = \sum_{k=1}^{n} a_k\lambda_k^p\mathbf{u}_k. \tag{22}$$

Hence, taking into account that the system of vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ is orthonormal and the matrix A is symmetric, we obtain the equality

$$(\mathbf{u}^T A^p,\quad A^p\mathbf{u}) = \sum_{k=1}^{n} a_k^2\lambda_k^{2p}, \tag{23}$$

where (x, y) is a scalar product of the vectors x and y. Consequently, for $a_1 \neq 0$ we may write

$$\frac{A^p\mathbf{u}}{\sqrt{(\mathbf{u}^T A^p,\ A^p\mathbf{u})}} = \frac{\sum_{k=1}^{n} a_k\lambda_k^p\mathbf{u}_k}{\sqrt{\sum_{k=1}^{n} a_k^2\lambda_k^{2p}}}$$

$$= \frac{a_1\lambda_1^p\left\{\mathbf{u}_1 + O\left[\left(\frac{\lambda_2}{\lambda_1}\right)^p\right]\right\}}{a_1\lambda_1^p\left\{1 + O\left[\left(\frac{\lambda_2}{\lambda_1}\right)^{2p}\right]\right\}} = \mathbf{u}_1 + O\left[\left(\frac{\lambda_2}{\lambda_1}\right)^p\right].$$

Thus, for sufficiently large values of p we have the expression for the eigenvector

$$\mathbf{u}_1 \cong \frac{A^p\mathbf{u}}{\sqrt{(\mathbf{u}^T A^p,\ A^p\mathbf{u})}}. \tag{24}$$

To determine λ_1, we write the relationship

$$\sqrt{(\mathbf{u}^T A^{p+1},\ A^{p+1}\mathbf{u})/(\mathbf{u}^T A^p,\ A^p\mathbf{u})}.$$

According to formula (23), we have

$$\sqrt{\frac{(\mathbf{u}^T A^{p+1},\ A^{p+1}\mathbf{u})}{(\mathbf{u}^T A^p,\ A^p\mathbf{u})}} - \frac{\sqrt{a_1^2\lambda_1^{2p+2} + a_2^2\lambda_2^{2p+2} + \ldots + a_n^2\lambda_n^{2p+2}}}{\sqrt{a_1^2\lambda_1^{2p} + a_2^2\lambda_2^{2p} + \ldots + a_n^2\lambda_n^{2p}}} = \frac{a_1\lambda_1^{p+1}\left\{1 + O\left[\left(\frac{\lambda_2}{\lambda_1}\right)^{2p+2}\right]\right\}}{a_1\lambda_1^{p}\left\{1 + O\left[\left(\frac{\lambda_2}{\lambda_1}\right)^{2p}\right]\right\}} = \lambda_1 + O\left[\left(\frac{\lambda_2}{\lambda_1}\right)^{2p}\right].$$

Therefore for sufficiently large values of p we have the approximate equality

$$\lambda_1 \cong \frac{\sqrt{(\mathbf{u}^T A^{p+1},\ A^{p+1}\mathbf{u})}}{\sqrt{(\mathbf{u}^T A^p,\ A^p\mathbf{u})}}. \tag{25}$$

The fulfilment of equality (25) with the preassigned accuracy serves as the criterion of finishing the iterative process.

If now, instead of the vector $\mathbf{u}$, we take the vector $\mathbf{u}^{(1)} = \mathbf{u} - (\mathbf{u},\ \mathbf{u}_1)\,\mathbf{u}_1 = a_2\mathbf{u}_2 + a_3\mathbf{u}_3 + \ldots + a_n\mathbf{u}_n$ orthogonal to the vector $\mathbf{u}_1$, then the repetition of the foregoing iterative process will determine the eigenvector $\mathbf{u}_2$ and eigenvalue λ_2. In such a way it is possible to determine all the eigenvectors and eigenvalues of the matrix A.

The determination of the eigenvalues and eigenvectors by this method is reduced to successive multiplications of the matrix A by a vector, which requires n^2 operations of multiplication. And if there are many zeros among the elements of the matrix A, as for instance in the case of a banded matrix, then the number of multiplications required for multiplying the matrix A by a vector is still reduced. Therefore for large n it is advisable to use this method for finding the eigenvalues and eigenvectors of the matrix A.

Remark. When computing the eigenvector $\mathbf{u}_k$, $k \geqslant 2$, the round-off errors are introduced by the components $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$, each time the vector $\mathbf{z}_m = A^m\mathbf{u}^{(k-1)}$ is computed, it is useful to define this vector more exactly by requiring that $\mathbf{z}_m$ be orthogonal to all the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k-1}$, that is, to take the vector $\tilde{\mathbf{z}} = \mathbf{z}_m - (\mathbf{z}_m, \mathbf{u}_1)\,\mathbf{u}_1 - \ldots - (\mathbf{z}_m, \mathbf{u}_{k-1})\,\mathbf{u}_{k-1}$ instead of the vector $\mathbf{z}_m$.

Sec. 10.2. SOLVING NONLINEAR EQUATIONS

1. Statement of the Problem. Consider the equation

$$F(\mathbf{x}) = 0, \tag{1}$$

where $u = F(\mathbf{x})$ is a continuous or differentiable operator mapping the domain Q of the m-dimensional vector space R onto the domain P of the same space.

The solution of equation (1), as a rule, cannot be found in the general form. Therefore, to determine this solution, we use the iterative methods based on reduction of equation (1) to the form

$$\mathbf{x} = \Phi(\mathbf{x}), \tag{2}$$

where $\Phi(\mathbf{x})$ is a compression operator defined in a neighbourhood G of the desired solution ($G \subset Q$). Starting from an arbitrary vector $\mathbf{x}^{(0)} \in G$, we determine in succession the vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}, \ldots$ by the formula

$$\mathbf{x}^{(n)} = \Phi(\mathbf{x}^{(n-1)}), \quad n = 1, 2, \ldots. \tag{3}$$

The limit of this sequence $\mathbf{x}^*$ is the solution of equation (2), and, by the same token, of equation (1) (see Sec. 3.4).

Thus the problem of solving equation (1) is divided into two stages: (1) determining the domain G in which equation (1) can be reduced to form (2), and (2) successive correction of the solution by formula (3).

The form of the function $\Phi(\mathbf{x})$ determines the method of solving equation (1).

The first stage of solving equation (1) for $m \geqslant 2$ is in practice not subject to formalization. The domain G, in which $\Phi(\mathbf{x})$ is a compression operator, is usually defined proceeding from the concrete form of equation (1) with regard to the physical meaning of the problem under consideration and complexity of computing the values of Φ on a computer.

2. Newton's Method for Solving One Equation. Consider the scalar equation

$$f(x) = 0 \tag{4}$$

in the supposition that the desired root c of this equation is isolated on the interval $[a, b]$, and the function $f(x)$ has

the first and second derivatives which do not change their sign on this interval. For the sake of definiteness, let us assume that both of these derivatives are positive on the interval $[a, b]$ (Fig. 45). Let us take the number $x_0 = b$ for the zero-order approximation to the desired root c, and let us denote the point with the coordinates $(x_0, f(x_0))$ by B_0. Draw through the point B_0 a tangent line to the graph of the function and take the abscissa $x_1 = x_0 - f(x_0)/f'(x_0)$ of the point of intersection of this tangent with the x-axis (see Fig. 45) for the first approximation to the desired root. Then draw a tangent line to the graph of the function through the point $B_1(x_1, f(x_1))$ and take the abscissa x_2 of the point of intersection of this tangent with the x-axis for the second approximation to the desired root. Continuing this process, we construct the sequence

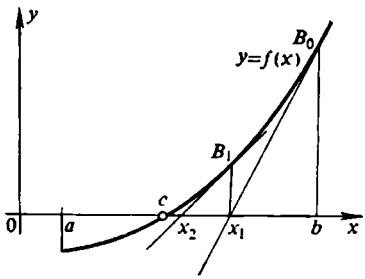


Fig. 45

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \ldots, \tag{5}$$

of approximate values of the desired root.

Let us show that our suppositions imply that sequence (5) has a limit.

It suffices to show that the sequence $\{x_n\}$ decreases monotonically and is bounded from below by the number c. Let $x_n > c$. Subtracting the equality

$$c = c - \frac{f(c)}{f'(x_n)}$$

from equality (5), we obtain the equalities

$$x_{n+1} - c = x_n - c - \frac{f(x_n) - f(c)}{f'(x_n)}$$

$$= (x_n - c)\left(1 - \frac{f'(\xi)}{f'(x_n)}\right), \quad c < \xi < x_n, \tag{6}$$

and since, by supposition, $f'(x) > 0$ and $f''(x) > 0$ for all $x \in [a, b]$, we have $0 < f'(\xi)/f'(x_n) < 1$. This shows that $0 < x_{n+1} - c < x_n - c$, i.e.

$$c < x_{n+1} < x_n. \tag{7}$$

Since the relationship $x_0 = b > c$ is fulfilled for $n = 0$, the method of complete induction implies that inequalities (7) hold true for all $n = 0, 1, \ldots$. Inequalities (7) show that the sequence $\{x_n\}$ decreases monotonically and is bounded from below by the number c. Therefore the sequence converges to a number d satisfying the condition $b > d \geqslant c$. But since the function $f(x)$ is continuous on $[a, b]$, passing to the limit in equality (5), we get the equality $d = d - f(d)/f'(d)$, which is equivalent to the equality $f(d) = 0$. By the supposition, the root c is isolated on $[a, b]$, therefore $c = d$, which proves the convergence of the sequence $\{x_n\}$ to the root c.

Let us now estimate the order of convergence. Here we assume that everywhere on the interval $[a, b]$ the inequalities $|f'(x)| \geqslant m > 0$ and $|f''(x)| \leqslant N$ are valid, and the interval $[a, b]$ is chosen so small that the following inequality holds true:

$$q = \frac{N}{2m}(b-a) < 1.$$

From (6) we obtain the relationships

$$x_{n+1} - c = \frac{f(c) - f(x_n) - f'(x_n)(c - x_n)}{f'(x_n)} = \frac{f''(\xi)(c-x_n)^2}{2f'(x_n)}, \quad a < \xi < b,$$

and, consequently,

$$|x_{n+1} - c| \leqslant \frac{N}{2m} |x_n - c|^2.$$

Applying successively this estimate for $n = 0, 1, \ldots$, we obtain the estimate

$$|x_{n+1} - c| \leqslant \left(\frac{N}{2m}\right)^{2^n - 1} |x_0 - c|^{2^n} < \left(\frac{N(b-a)}{2m}\right)^{2^n - 1} (b-a) = q^{2^n - 1}(b-a) \tag{8}$$

showing a very rapid convergence of the sequence $\{x_n\}$ to c.

3. Newton's Method for Solving Systems of Equations. Let us write equation (1) in the form of a system of equations

$$\begin{aligned} f_1(x_1, \ldots, x_m) &= 0, \\ \ldots\ldots\ldots\ldots & \\ f_m(x_1, \ldots, x_m) &= 0, \end{aligned} \tag{9}$$

and let us assume that in a neighbourhood $G \in R$ system (9) has the unique solution $(c_1, c_2, \ldots, c_m)$. Suppose that all the functions $f_j(x_1, \ldots, x_m)$ are at least doubly differentiable in the domain $\overline{G}$, the following inequality being valid in this domain:

$$\frac{D(f_1, \ldots, f_m)}{D(x_1, \ldots, x_m)} = \begin{vmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \cdots & \cdots & \cdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_m} \end{vmatrix} \neq 0. \tag{10}$$

Let the vector $(x_1^{(n)}, \ldots, x_m^{(n)})$ belong to G and be the nth approximation to the solution $(c_1, \ldots, c_m)$. Then it is possible to write the equalities

$$\begin{aligned} 0 = f_l(c_1, \ldots, c_m) = f_l(x_1^{(n)}, \ldots, x_m^{(n)}) \\ + \sum_{k=1}^{n} \frac{\partial f_l(x_1^{(n)}, \ldots, x_m^{(n)})}{\partial x_k}(c_k - x_k^{(n)}) \\ + \frac{1}{2} \sum_{k,\, l=0}^{m} \left[\frac{\partial^2 f(\xi_1, \ldots, \xi_m)}{\partial x_k \partial x_l}(c_k - x_k^{(n)})(c_l - x_l^{(n)}) \right]. \end{aligned} \tag{11}$$

Since the last sum has the order $O(\| c - x^{(n)} \|^2)$, then if the norm of the difference $\| c - x^{(n)} \|$ is sufficiently small, it may be rejected. The system of equations obtained from (11) can be applied to determine the following approximation:

$$\sum_{k=1}^{n} \frac{\partial f_l(x_1^{(n)}, \ldots, x_m^{(n)})}{\partial x_k}(x_k^{(n+1)} - x_k^{(n)}) = -f_l(x_1^{(n)}, \ldots, x_m^{(n)}),$$

$$l = 1, 2, \ldots, m. \tag{12}$$

By virtue of condition (10), system (12) has the unique solution $(x_1^{(n+1)}, \ldots, x_m^{(n+1)})$.

Thus, it is possible to construct a sequence of vectors

$$(x_1^{(0)}, \ldots, x_m^{(0)}), \ldots, (x_1^{(n)}, \ldots, x_m^{(n)}) \qquad (13)$$

which will converge to the desired solution $(c_1, \ldots, c_m)$ provided the norm $\| \mathbf{x}^{(0)} - \mathbf{c} \|$ is sufficiently small.

Such an iterative process is called *Newton's method.*

The order of convergence established by estimate (8) is preserved when Newton's method is applied for solving systems of equations. Therefore this method is widely used for solving systems of equations of form (9) on a computer despite the complications due to the necessity to compute the values of partial derivatives and to solve a system of linear equations (12).

Sec. 10.3. NUMERICAL METHODS OF SOLVING DIFFERENTIAL EQUATIONS

1. Euler's Method of Solving Cauchy's Problem. Differential equations are frequently used in applied problems. If a problem is reduced to solving a system of ordinary linear differential equations with constant coefficients as, for instance, the majority of problems in the theory of electric circuits, then its solution can be found in explicit form. But if the differential equations have variable coefficients or are nonlinear, then their solution has to be found numerically. Computers considerably facilitate the solution of differential equations.

Consider the differential equation

$$y' = f(x, y) \qquad (1)$$

in the supposition that the function $f(x, y)$ is differentiable in a neighbourhood of the point (x_0, y_0). Cauchy's problem for the differential equation (1) is formulated in the following way: *Find the solution $y(x)$ of equation (1) satisfying the condition $y(x_0) = y_0$.*

Let us assume that the solution $y(x)$ of equation (1) is known at the point x_n and it is required to find $y(x_n + h)$.

From the obvious equality

$$y(x_n+h)=y(x_n)+\int_{x_n}^{x_n+h} y'(x)\,dx,$$

taking into consideration equation (1), we get the equality

$$y(x_n+h)=y(x_n)+\int_{x_n}^{x_n+h} f(x,\ y(x))\,dx. \tag{2}$$

By Taylor's formula applied to expression (2), we find the equality

$$y(x_n+h)=y(x_n)+hf(x_n,\ y(x_n))+O(h^2).$$

Rejecting the terms of order $O(h^2)$ in this expression and setting $x_{n+1}=x_n+h$, $y_n=y(x_n)$, $y_{n+1}=y(x_{n+1})$, we obtain Euler's formula

$$y_{n+1}=y_n+hf(x_n,\ y_n),\quad n=0,\ 1,\ \ldots. \tag{3}$$

The error of formula (3) will have the order $O(h^2)$.

In order to obtain a more accurate computational formula than formula (3), let us evaluate the integral on the right-hand side of equality (2) using the trapezoidal formula (see Sec. 9.2). We have

$$y(x_n+h)=y(x_n)+\frac{h}{2}[f(x_n,\ y(x_n))+f(x_{n+1},\ y(x_{n+1}))].$$

By Taylor's formula, the equality $f(x_{n+1},\ y(x_{n+1}))=f(x_{n+1},\ y_n+hf(x_n,\ y_n))+O(h^2)$ holds true. Rejecting the terms of order $O(h^3)$ in the preceding expression and setting $y_{n+1}^*=y_n+hf(x_n,\ y_n)$, we get the following formulas:

$$\begin{aligned} y_{n+1}^* &= y_n+hf(x_n,\ y_n),\\ y_{n+1} &= y_n+\frac{h}{2}[f(x_n,\ y_n)+f(x_{n+1},\ y_{n+1}^*)], \end{aligned} \tag{4}$$

whose error has the value of order $O(h^3)$. Formulas (4) are called the *Euler-Cauchy formulas.*

2. Analysis of Total Error. Let $\bar{y}(x)$ be the true solution of Cauchy's problem for equation (1). Setting

$$\eta_0=0,\quad \eta_n=\bar{y}(x_n)-y_n,\quad n=1,\ 2,\ \ldots, \tag{5}$$

we pass from equation (1) to the integral equation (2)

$$\bar{y}(x_n+h)=\bar{y}(x_n)+\int\limits_{x_n}^{x_n+h} f(x,\ \bar{y}(x))\,dx.$$

Let us write equality (3) in the form

$$y_{n+1}=y_n+\int\limits_{x_n}^{x_n+h} f(x_n,\ y_n)\,dx$$

and subtract it from the preceding equality. Bearing in mind (5), we obtain the equality

$$\eta_{n+1}=\eta_n+\int\limits_{x_n}^{x_n+h}\left[\frac{\partial f(x_n,\ y_n)}{\partial y}\eta_n\right.$$

$$\left.+\left(\frac{\partial f(x_n,\ y_n)}{\partial x}+\frac{\partial f(x_n,\ y_n)}{\partial y}f(x_n,\ y_n)\right)(x-x_n)\right]dx+O(h^3).$$

Rejecting the terms of order $O(h^3)$, we obtain the difference equation

$$\eta_{n+1}=\left(1+h\frac{\partial f(x_n,\ y_n)}{\partial y}\right)\eta_n$$

$$+\frac{h^2}{2}\left(\frac{\partial f(x_n,\ y_n)}{\partial x}+\frac{\partial f(x_n,\ y_n)}{\partial y}f(x_n,\ y_n)\right). \quad (6)$$

Suppose we have to find the solution of equation (1) on the interval $[x_0,\ x_0+T]$. Let us introduce the notation $[T/h]=N$ and suppose that for all $n=0,\ 1,\ \ldots,\ N$ the following inequalities are fulfilled:

$$a\leqslant\frac{\partial f(x_n,\ y_n)}{\partial y}\leqslant b,$$

$$\left|\frac{\partial f(x_n,\ y_n)}{\partial x}+\frac{\partial f(x_n,\ y_n)}{\partial y}f(x_n,\ y_n)\right|\leqslant c.$$

We shall also assume that $1+ha\geqslant 0$, $|b|h<1$. Consequently, the following inequalities hold true:

$$0\leqslant 1+ha\leqslant 1+h\frac{\partial f(x_n,\ y_n)}{\partial y}\leqslant 1+hb.$$

Applying them to equality (6), we get the inequality

$$|\eta_{n+1}| \leqslant (1+hb)|\eta_n| + \frac{1}{2}ch^2, \quad \eta_0 = 0. \tag{7}$$

Let β_n be the solution of the difference equation

$$\beta_{n+1} = (1+hb)\beta_n + \frac{1}{2}ch^2, \quad \beta_0 = 0. \tag{8}$$

Comparing expressions (7) and (8) for $n = 0, 1, \ldots, N$, we make sure that

$$|\eta_n| \leqslant \beta_n, \quad n = 0, 1, \ldots, N. \tag{9}$$

The solution β_n of equation (8) can be written in the form

$$\beta_n = \frac{1}{2}\frac{ch}{b}[(1+hb)^n - 1] \quad \text{for } b \neq 0,$$

$$\beta_n = \frac{1}{2}ch^2 n \quad \text{for } b = 0.$$

Taking into account the inequality $h \leqslant T/N$, we obtain that for all $n = 1, \ldots, N$ the following inequalities are valid:

$$|\eta_n| \leqslant \frac{ch}{2b}(e^{Tb} - 1) \quad \text{for } b \neq 0,$$

$$|\eta_n| \leqslant \frac{cTh}{2} \quad \text{for } b = 0. \tag{10}$$

Inequalities (10) show that the total error in integrating equation (1) by Euler's method is a quantity of order $O(h)$.

3. The Runge–Kutta Methods. Let us suppose that the function $f(x, y)$ has continuous partial derivatives up to the mth order inclusively, then the solution $y(x)$ of Cauchy's problem for equation (1) will possess continuous derivatives up to the $(m+1)$th order inclusively, and if the value of $y(x)$ for $x = x_n$ is known, $y(x_n) = y_n$, then the following equality is valid:

$$y(x_n + h) = y(x_n) + y'(x_n)h + \frac{1}{2}y''(x_n)h^2 + \ldots + \frac{y^{(m)}(x_n)}{m!}h^m + \frac{y^{(m+1)}(\xi)}{(m+1)!}h^{m+1}, \quad x_n < \xi < x_n + h. \tag{11}$$

The values of the derivatives entering into this equality are computed from equation (1) by successive differentiation:

$$\begin{aligned} y'(x_n) &= f(x_n, y_n), \\ y''(x_n) &= f'_x(x_n, y_n) + f'_y(x_n, y_n) f(x_n, y_n), \\ y'''(x_n) &= f''_{x^2}(x_n, y_n) + 2f(x_n, y_n) f''_{xy}(x_n, y_n) \\ &\quad + f^2(x_n, y_n) f''_{y^2}(x_n, y_n) + (f'_x(x_n, y_n) \\ &\quad + f(x_n, y_n) f'_y(x_n, y_n)) f'_y(x_n, y_n), \ \ldots . \end{aligned} \tag{12}$$

Substituting the values of $y'(x_n)$, $y''(x_n)$, ... determined by expressions (12) into relationship (11), it is possible to compute the value of $y(x_n + h) = y_{n+1}$. But such calculation requires computations by formulas (12) whose complexity increases rapidly with an increase in the order of derivatives. In order to reduce the computational work, G. Runge suggested to seek the value of $y(x_n + h)$ in the form

$$y(x_n + h) = y(x_n) + M_s(h) + O(h^{s+1}), \tag{13}$$

where

$$\begin{aligned} M_s(h) &= p_1 k_1(h) + p_2 k_2(h) + \ldots + p_s k_s(h), \\ k_1(h) &= hf(x_n, y_n), \\ k_2(h) &= hf(x_n + \alpha_2 h, y_n + \beta_{21} k_1(h)), \\ \ldots, k_s(h) &= hf(x_n + \alpha_s h, y_n + \beta_{s1} k_1(h) \\ &\quad + \ldots + \beta_{s,s-1} k_{s-1}(h)), \end{aligned}$$

$p_1, p_2, \ldots, p_s, \alpha_2, \ldots, \alpha_s, \beta_{21}, \ldots, \beta_{s,s-1}$ are certain parameters. Formula (3) is obtained as a particular case of formula (13) for $s = 1$, and formula (4) for $s = 2$. Consider the question concerning the choice of the parameters p_m, α_m, β_{mi}. For the sake of simplicity, we shall confine ourselves to the case $s = 3$. Let us introduce the notation

$$\begin{aligned} \varphi(h) &= y(x_n + h) - y(x_n) \\ &\quad - p_1 k_1(h) - p_2 k_2(h) - p_3 k_3(h). \end{aligned} \tag{14}$$

From expression (13) it follows that

$$\varphi(0) = \varphi'(0) = \varphi''(0) = \varphi'''(0) = 0. \tag{15}$$

Taking into account relationships (12), from equality (14) we find:

$$\varphi(0) = 0,$$

$$\varphi'(0) = (1 - p_1 - p_2 - p_3) f,$$

$$\varphi''(0) = (1 - 2p_2\alpha_2 - 2p_3\alpha_3) f'_x + [1 - 2p_2\beta_{21} - 2p_3(\beta_{31} + \beta_{32})] f'_y f,$$

$$\varphi'''(0) = [1 - 3(p_2\alpha_2^2 + p_3\alpha_3^2)] f''_{x^2} + 2[1 - 3(p_2\alpha_2\beta_{21} + p_3\alpha_3\beta_{31} + p_3\alpha_3\beta_{32})] f''_{xy} + \{1 - 3[p_2\beta_{21}^2 + p_3(\beta_{31} + \beta_{32})^2]\} f''_{y^2} f^2 + (1 - 6p_3\alpha_2\beta_{32})(f')^2 f.$$

Conditions (15) will be fulfilled if the following equalities are valid:

$$\begin{gathered} p_1 + p_2 + p_3 = 1, \quad p_2\alpha_2 + p_3\alpha_3 = 1/2, \\ p_2\alpha_2^2 + p_3\alpha_3^2 = 1/3, \\ \beta_{21} = \alpha_2, \quad \beta_{31} + \beta_{32} = \alpha_3, \quad p_3\alpha_2\beta_{32} = 1/6. \end{gathered} \tag{16}$$

This system of six equations in eight unknowns has infinitely many solutions. The most commonly used solution

$$\begin{gathered} p_1 = 1/6, \quad p_2 = 4/6, \quad p_3 = 1/6, \\ \alpha_2 = 1/2, \quad \beta_{21} = 1/2, \\ \alpha_3 = 1, \quad \beta_{31} = -1, \quad \beta_{32} = 2 \end{gathered}$$

yields the computational formulas:

$$\begin{gathered} k_1(h) = hf(x_n, y_n), \\ k_2(h) = hf\left(x_n + \frac{h}{2}, \; y_n + \frac{1}{2}k_1(h)\right), \\ k_3(h) = hf[x_n + h, \; y_n - k_1(h) + 2k_2(h)], \\ y_{n+1} = y_n + \frac{1}{6}[k_1(h) + 4k_2(h) + k_3(h)], \quad n = 0, 1, \ldots . \end{gathered} \tag{17}$$

For $s = 4$ the following computational formulas are obtained:

$$\begin{gathered} k_1(h) = hf(x_n, y_n), \\ k_2(h) = hf\left(x_n + \frac{h}{2}, \; y_n + \frac{1}{2}k_1(h)\right), \end{gathered} \tag{18}$$

$$k_3(h) = hf\left(x_n + \frac{h}{2},\ y_n + \frac{k_2(h)}{2}\right),$$

$$k_4(h) = hf[x_n + h,\ y_n + k_3(h)],$$

$$y_{n+1} = y_n - \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad n = 0, 1, 2, \ldots.$$

The estimate of the error obtained for each step of integration of equation (1) by the Runge-Kutta method, according to formula (13), has the order $O(h^{s+1})$. The total error can be estimated in the same way as it was done for Euler's method. Since in this case the relevant computations become more complicated, we are not going to dwell on them here.

4. Adams' Method. Another method for determining partial sums of series (11) without finding the values of derivatives by formulas (12) was suggested by J. Adams. Let there be known the solution $y(x)$ of equation (1) at the points x_0, $x_1 = x_0 + h$, ..., $x_n = x_0 + nh$; $y(x_0) = y_0$, ..., $y(x_n) = y_n$. According to expansion (11), we write the equality

$$y(x_n + h) = y_n + A_1 h + A_2 \frac{h^2}{2!} + A_3 \frac{h^3}{3!} + A_4 \frac{h^4}{4!} + A_5 \frac{h^5}{5!} + \ldots,$$

where

$$A_i = y^{(i)}(x_n), \quad i = 1, 2, \ldots. \tag{19}$$

Differentiating expression (19) with respect to h, we obtain the equality

$$y'(x_n + h) = A_1 + A_2 h + A_3 \frac{h^2}{2!} + A_4 \frac{h^3}{3!} + A_5 \frac{h^4}{4!} + \ldots. \tag{20}$$

Setting $y'(x_i)\, h = \eta_i$, we find from expression (20):

$$\eta_n = A_1 h,$$

$$\eta_{n-1} = A_1 h - A_2 h^2 + A_3 \frac{h^3}{2!} - A_4 \frac{h^4}{3!} + A_5 \frac{h^5}{4!} - \ldots,$$

$$\eta_{n-2} = A_1 h - 2A_2 h^2 + 4A_3 \frac{h^3}{2!} - 8A_4 \frac{h^4}{3!} + 16A_5 \frac{h^5}{4!} - \ldots,$$

$$\eta_{n-3} = A_1 h - 3A_2 h^2 + 9A_3 \frac{h^3}{2!} - 27A_4 \frac{h^4}{3!} + 81A_5 \frac{h^5}{4!} - \ldots.$$

These expressions enable us to form the finite differences:

$$\Delta\eta_{n-1} = A_2h^2 - A_3\frac{h^3}{2!} + A_4\frac{h^4}{3!} - A_5\frac{h^5}{4!} + \ldots,$$

$$\Delta^2\eta_{n-2} = 2A_3\frac{h^3}{2!} - 6A_4\frac{h^4}{3!} + 14A_5\frac{h^5}{4!} - \ldots,$$

$$\Delta^3\eta_{n-3} = 6A_4\frac{h^4}{3!} - 36A_5\frac{h^5}{4!} + \ldots.$$

Hence we find the values of the coefficients of series (19) in form of the equalities

$$A_2h^2 = \Delta\eta_{n-1} + \frac{1}{2}\Delta^2\eta_{n-2} + \frac{1}{3}\Delta^3\eta_{n-3} + A_5\frac{h^5}{4!} + \ldots,$$

$$A_3h^3 = \Delta^2\eta_{n-2} + \Delta^3\eta_{n-3} + \frac{11}{12}A_5h^5 + \ldots.$$

$$A_4h^4 = \Delta^3\eta_{n-3} + \frac{3}{2}A_5h^5 + \ldots.$$

Substituting the found values of A_1, A_2, A_3, A_4 into relationship (19), we get

$$y(x_n+h) = y_n + \eta_n + \frac{1}{2}\Delta\eta_{n-1} + \frac{5}{12}\Delta^2\eta_{n-2} + \frac{3}{8}\Delta^3\eta_{n-3} + \frac{251}{720}A_5h^5 + \ldots.$$

Rejecting the remainder term, we get the computational formula

$$y_{n+1} = y_n + \eta_n + \frac{1}{2}\Delta\eta_{n-1} + \frac{5}{12}\Delta^2\eta_{n-2} + \frac{3}{8}\Delta^3\eta_{n-3}. \quad (21)$$

Formula (21) requires that the values of y_0, y_1, y_2, y_3 be known before it is applied. The values of y_1, y_2, y_3 can be determined by the Runge-Kutta formula (18) having the same order of accuracy as formula (21). The calculation of the differences entering into formula (21) requires that only one set of values of $f(x, y)$ be computed, which reduces the number of operations necessary for solving equation (1).

The foregoing methods of constructing the formulas for numerical integration of one differential equation are also applicable to the case of a system of equations:

$$y_i' = f_i(x, y_1, \ldots, y_n), \quad i = 1, \ldots, m. \quad (22)$$

The formulas for computing the derivatives $y_i^{(k)}$ in this case are obtained in the form

$$\begin{aligned} y_i' &= f_i, \\ y_i'' &= \frac{\partial f_i}{\partial x} + \sum_{j=1}^{m} \frac{\partial f_i}{\partial y_i} f_j, \\ y_i''' &= \frac{\partial^2 f_i}{\partial x^2} + 2\sum_{j=1}^{m} \frac{\partial^2 f_i}{\partial x\, \partial y_j} f_j + \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{\partial^2 f_i}{\partial y_j\, \partial y_k} f_j f_k \\ &\quad + \sum_{j=1}^{m} \frac{\partial f_i}{\partial y_j}\left(\frac{\partial f_j}{\partial x} + \sum_{l=1}^{m} \frac{\partial f_j}{\partial y_l} f_l\right). \end{aligned} \tag{23}$$

Computational formulas of form (18) or (21) are applied to each of the equations of system (22) separately.

Sec. 10.4. NET-POINT METHOD

The finite-difference approximations of derivatives are also useful for solving partial differential equations. For the sake of simplicity, let us confine ourselves to the case of two variables and consider the equation

$$Lu \equiv A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x\, \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu = G, \tag{1}$$

whose coefficients $A, \ldots, F$ and the constant term G are bounded functions of x and y in the domain Q.

The following problem is set: *Find the function u satisfying equation* (1) *in the domain Q and the conditions*

$$Hu \equiv a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} + cu = g \tag{2}$$

on the boundary Γ of the domain Q (or on the portion of this boundary).

For the numerical solution of this problem, with the aid of the straight lines $x_i = ih$ and $y_k = k\tau$, $i, k = 0, 1, \ldots$, we construct a rectangular net S_h covering the domain Q.

The function $v_h(x, y)$ defined at the points $M_{ik} = (x_i, y_k)$ of the net S_h is called the *net function*, setting $v_{ik} = v_h(x_i, y_k)$. The problem of the numerical solution of equation (1) consists in determining a net function u_h such that its deviation from the solution u of equation (1) at all the net points M_{ik} belonging to the domain Q does not exceed the given error. In order to solve this problem, in most cases the derivatives in the operator Lu are replaced by difference relationships with the help of the formulas:

$$\frac{\partial u}{\partial x} = \frac{u_{i+1,k} - u_{ik}}{h}, \quad \frac{\partial u}{\partial y} = \frac{u_{i,k+1} - u_{ik}}{\tau},$$

$$\frac{\partial^2 u}{\partial x \partial y} = \frac{u_{i+1,k+1} - u_{i+1,k-1} - u_{i-1,k+1} + u_{i-1,k-1}}{4h\tau}, \qquad (3)$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{u_{i+1,k} - 2u_{ik} + u_{i-1,k}}{h^2}, \quad \frac{\partial^2 u}{\partial y^2} = \frac{u_{i,k+1} - 2u_{ik} + u_{i,k-1}}{\tau^2}.$$

On performing such operations for all the points M_{ik} of the net S_h from the domain Q, we obtain the system of difference equations

$$L_h u_{ik} \equiv \sum_{j=-1}^{1} \sum_{l=-1}^{1} p_{i+j,k+l} u_{i+j,k+l} = G_{ik}. \qquad (4)$$

Similarly, from the boundary conditions (2) it is possible to get the equations

$$H_h u_{ik} = \sum_{k=0}^{1} \sum_{j=0}^{1} r_{i+l,k+j} u_{i+l,k+j} = g_{ik} \qquad (5)$$

for the points M_{ik} of the net S_h situated near or on the boundary Γ of the domain Q. The combination of equations (4) and (5) is referred to as the *difference scheme*, its solution u_h is termed the *net-point solution* of equation (1), and the method of determining the numerical solution u_h of equation (1) as the solution of the system of equations (4) and (5) is called the *net-point method*. Owing to the large number of unknowns which we have to deal with in the net-point method, the analysis of total error is still of greater importance than in the case of ordinary differential equations. Its leading features are the notions of approximation and stability. The difference operator $L_h u$ is said to be an *ap-*

proximation to the differential operator Lu of order p with respect to h and q with respect to τ if the following relationships are fulfilled:

$$| Lu(x_i, y_k) - L_h u_{ik} | = O(h^p + \tau^q), \quad p, q \geqslant 1. \tag{6}$$

A difference scheme is said to be *stable* if the system of equations (4) and (5) has a unique solution and there are positive numbers c_1 and c_2 independent of h, τ, G, and g such that

$$\| u_h \| \leqslant c_1 \| G_h \| + c_2 \| g_h \|. \tag{7}$$

It is known that if a difference scheme is an approximation and is stable, then for h and τ tending to zero, the net-point solution u_h tends to the desired solution u of equation (1). But if the difference scheme is unstable, then the computational error may become too large and it is impossible to apply such a scheme.

We shall carry out the analysis of questions concerning the application of the net-point method for solving concrete equations for the wave equation. During the process of solving this equation the peculiar features characteristic for the net-point method are manifested to a sufficient extent. Let us take the equation

$$Lu \cong \frac{\partial^2 u}{\partial t^2} - a^2 \frac{\partial^2 u}{\partial x^2} = f(x, t) \tag{8}$$

and seek its solution in the domain D: $0 \leqslant x \leqslant 1, 0 < t \leqslant T$, for the conditions

$$u(x, 0) = \varphi_1(x), \quad \frac{\partial u(x, 0)}{\partial t} = \varphi_2(x) \text{ (initial conditions)}, \tag{9}$$

$$u(0, t) = \psi_1(t), \quad u(1, t) = \psi_2(t) \text{ (boundary conditions)}. \tag{10}$$

In the domain D let us introduce the net S_h: $x_i = ih$, $i = 0, \ldots, m$; $t_k = k\tau$, $k = 0, \ldots, r$. Taking advantage of formulas (3), we find the difference equations

$$\frac{u_{i, k+1} - 2u_{ik} + u_{i, k-1}}{\tau^2} - a^2 \frac{u_{i+1, k} - 2u_{ik} + u_{i-1, k}}{h^2} = f_{ik}. \tag{11}$$

It is not difficult to show that if $u(x, t)$ has continuous derivatives up to the fourth order inclusively, then the

net-point equation (11) is an approximation of equation (8) of order $O(h^2 + \tau^2)$. The boundary conditions (10) and the first of the initial conditions (9) on the net S_h are fulfilled exactly. The second initial condition (9) is approximated with the aid of the hypothetical layer $t = -\tau$. We have

$$\frac{u(x, 0) - u(x, -\tau)}{\tau} = \frac{\partial u(x, 0)}{\partial t} - \frac{1}{2}\frac{\partial^2 u(x, 0)}{\partial t^2} + O(\tau^2).$$

Hence, taking into consideration equation (8) and the first condition of (9), we find the difference equation

$$\frac{\varphi_{1i} - u_{i,-1}}{\tau} + \frac{1}{2}\tau\left(\frac{a^2(\varphi_{1,i+1} - 2\varphi_{1,i} + \varphi_{1,i-1})}{h^2} + f_{i0}\right) = \varphi_{2i} \quad (12)$$

which approximates the second of the initial conditions with an error of order $O(\tau^2 + h^2)$. Denoting $\sigma = a^2\tau^2/h^2$, we can write the difference scheme for solving equation (8) in the form

$$u_{i,k+1} = \sigma u_{i-1,k} + 2(1-\sigma)u_{ik} + \sigma u_{i+1,k} - u_{i,k-1} + \tau^2 f_{ik},$$
$$u_{i,-1} = \frac{\sigma}{2}\varphi_{1,i-1} + (1-\sigma)\varphi_{1i} + \frac{\sigma}{2}\varphi_{1,i+1} - \tau\varphi_{2i} + \frac{\tau^2}{2}f_{i0}, \quad (13)$$
$$u_{i0} = \varphi_{1i}, \quad u_{0k} = \psi_{1k}, \quad u_{mk} = \psi_{2k}.$$

We know $u_{i,-1}$ and u_{i0} for $i = 0, 1, \ldots, m$. Therefore, setting $k = 0$, from the first relations (13) we find u_{i1}, $i = 1, \ldots, m-1$, then, setting $k = 1$, we find u_{i2} and so on until the net function u_h is defined completely.

The difference schemes in which, like in scheme (13), we can determine all values of the net-point solution by successive substitution of already known values of the net-point solution into difference equations are called *explicit*, otherwise they are termed *implicit*.

Confining ourselves, for simplicity, to the case of a homogeneous equation [$f(x, t) = 0$] and homogeneous boundary conditions [$\psi_1(t) = \psi_2(t) = 0$], we investigate the difference scheme (13) for stability. To this end, following Fourier's method, we seek the particular solution $u_{ik} = X_i T_k$. Substituting it into the first equation of (13) and separating the variables, we get

$$\frac{T_{k+1} - 2T_k + T_{k-1}}{\sigma T_k} = \frac{X_{i+1} - 2X_i + X_{i-1}}{X_i} = \lambda, \quad (14)$$

where λ is the parameter of separation. Equation (14) falls into two equations

$$X_{l+1} - (2 + \lambda) X_l + X_{l-1} = 0, \quad X_0 = X_m = 0, \tag{15}$$

$$T_{k+1} - (2 + \lambda\sigma) T_k + T_{k-1} = 0. \tag{16}$$

If we seek the solution of problem (15) in the form $X_l = \sin \frac{\pi l l}{m}$, then we have $\sin \frac{\pi l i}{m} \left(2 \cos \frac{\pi l}{m} - 2 - \lambda\right) = 0$, and, consequently, problem (15) has a nontrivial solution if

$$\lambda = 2\left(\cos \frac{\pi l}{m} - 1\right) = -4 \sin^2 \frac{\pi l}{2m}, \quad l = 1, \ldots, m-1. \tag{17}$$

The solution of equation (16) has the form $T_k = c_1 \mu_1^k + c_2 \mu_2^k$, where μ_1 and μ_2 are the roots of the characteristic equation

$$\mu^2 - 2\left(1 - 2\sigma \sin^2 \frac{\pi l}{2m}\right)\mu + 1 = 0. \tag{18}$$

Let $\sigma \leqslant 1$, then equation (18) has two complex conjugate roots μ_1 and μ_2. Since $\mu_1 \mu_2 = 1$, we obtain

$$\mu_{1,2} = \cos \omega_l \pm i \sin \omega_l, \tag{19}$$

where

$$\sin \frac{\omega_l}{2} = \sqrt{\sigma} \sin \frac{\pi l}{2m}. \tag{20}$$

Bearing in mind that $\mu_{1,2}^k = \cos k\omega_l \pm i \sin k\omega_l$, we see that T_k has the form

$$T_k = a_l \cos k\omega_l + b_l \sin k\omega_l,$$

and the general solution of equation (13) is defined by the equality

$$u_{ik} = \sum_{l=1}^{m-1} (a_l \cos k\omega_l + b_l \sin k\omega_l) \sin \frac{\pi l i}{m}. \tag{21}$$

To determine the coefficients a_l and b_l, let us take advantage of the initial conditions (9). We get the equations

$$\sum_{l=1}^{m-1} a_l \sin \frac{\pi l i}{m} = \varphi_{1i}, \quad \sum_{l=1}^{m-1} b_l \frac{\omega_l}{\tau} \sin \frac{\pi l i}{m} = \varphi_{2i}. \tag{22}$$

The numbers a_l and b_l are found from these equations as the coefficients of the discrete Fourier transform (see Part 1,

Chapter 10) of the functions $\varphi_1(x)$ and $\varphi_2(x)$. Writing Parseval's equalities for relationships (22), we get

$$\sum_{l=1}^{m-1} a_l^2 = \frac{2}{m} \sum_{i=1}^{m-1} \varphi_{1i}^2, \quad \sum_{l=1}^{m-1} b_l^2 \left(\frac{\omega_l}{\tau}\right)^2 = \frac{2}{m} \sum_{i=1}^{m-1} \varphi_{2i}^2. \quad (23)$$

From relationships (20) we obtain a chain of inequalities

$$\frac{\omega_l}{2} \geqslant \sin\frac{\omega_l}{2} = \sqrt{\sigma} \sin\frac{l\pi}{2m} \geqslant \frac{2}{\pi}\sqrt{\sigma}\frac{l\pi}{2m} = \frac{a\tau l}{mh} = a\tau l \geqslant a\tau.$$

Taking into account this chain, we get from equalities (23) the estimate

$$\sum_{l=1}^{m-1} b_l^2 \leqslant \frac{1}{2a^2 m} \sum_{i=1}^{m-1} \varphi_{2i}^2. \quad (24)$$

The norm $\| g \|_2$ of the net function $g(x)$ is defined by the equality

$$\| g \|_2 = \left(\frac{1}{m} \sum_{i=1}^{m-1} g_i^2\right)^{1/2}.$$

Squaring both sides of equality (21) and summing with respect to l, we obtain from Parseval's equality the relationships[1]

$$\sum_{i=1}^{m-1} u_{ik}^2 = \frac{m}{2} \sum_{l=1}^{m-1} (a_l \cos k\omega_l + b_l \sin k\omega_l)^2 \leqslant \frac{m}{2} \sum_{l=1}^{m-1} (a_l^2 + b_l^2),$$

wherefrom, taking advantage of (23) and (24), we get the estimate

$$\| u(x,\ k\tau) \|_2^2 \leqslant \| \varphi_1 \|_2^2 + \frac{1}{4a^2} \| \varphi_2 \|_2^2. \quad (25)$$

This estimate proves the stability of the difference scheme (13).

Consider now the case $\sigma > 1$. In this case for large m and l sufficiently close to m equation (18) has distinct real roots μ_1 and μ_2, and since $\mu_1\mu_2 = 1$, one of the numbers μ_1, μ_2, say $| \mu_1 | > 1$. Then the particular solution $u_{lk} = \mu_1^k \sin\frac{\pi l i}{m}$ highly increases with an increase in k, and, consequently, the net-point solution u_k increases without bound for τ tending to zero, which indicates that the difference scheme (13) is unstable for $\sigma > 1$.

INDEX

Abelian group, 153
Accessory conditions, 227
Action, 257
Affine transformation, 125
Alternance,
  Chebyshev's, 298
  Vallée-Poussin's, 298
Application
  of contraction mapping principle to solution of equations, 140
  of Galerkin's method directly to solving boundary-value problems, 270
  of Hilbert-Schmidt theorem, 190-202
  of interpolation to problems of numerical differentiation and integration, 304-338
  of Ostrogradsky-Gauss formula to a special kind of the vector field, 86
  of Schauder's theorem, 163-172
  of Tailor's formula to the difference of equations, 228
Arzela's theorem, 148
Axiom of the metric, 122

Banach-Steinhaus theorem, 307
Bernstein's theorem, 298
Bolzano-Weierstrass theorem, 145
Brachistochrone, 203
Brouwer's theorem, 164

Cauchy's sequence, 133
Cauchy-Buniakowski inequality, 117
Chebyshev polynomials of the first kind, 301
Chebyshev's theorem, 29
Circulation of vector field, 51
Compact sets, 144-152
Completely ordered set, 168
Computation of curl in Cartesian coordinates, 54
Computational error, 285
Computing the norm of a self-adjoint operator, 183
Condition(s)
  accessory, 227
  field potentiality, 67-71
  isoperimetric, 204
  natural boundary, 247-248
  transversality, 249
Construction problems, 290
Continuity equation, 104
Continuous operator, 131, 156
Contour, 11
Contraction mapping principle, 139
Convergence in norm, 154
Coordinate
  axes, 15
  basis, 15
  line, 15
  surface, 15
  system, 15
Cubic spline, 317
Curl, 53
  computation of, in Cartesian

Curl,
coordinates, 54
in an orthogonal curvilinear coordinate system, 58
Curve(s),
orientation of, 10
negative, 10
positive, 10
oriented, 10
piecewise smooth, 11
smooth, 9

Difference scheme(s),
explicit, 364
implicit, 364
stalbe, 363
Differential operations of the second order, 84
Direct methods in the calculus of variations, 263
Directional derivative, 21
Distance between the elements, 122
Divergence, 45-51
computing of, in Cartesian coordinates, 46-47
in orthogonal curvilinear coordinates, 50-51
of a field at a point, 46
notion of, 45
properties of, 47-48
Divided difference(s),
first-order, 311
n-order, 311
second-order, 311
Domain,
multiply connected, 59
simple, 32
simply connected, 59
superficially simply connected, 60

Element of best approximation, 289
Equation(s),
continuity, 104
Euler's, 215
particular cases of, 217-222
Fredholm's integral, 174
Fredholm symmetric integral, 194
heat, 108
Laplace's, 88
of mathematical physics, 104-110
membrane motion, 260
Poisson's, 94, 254
of string vibrations, 259
Error(s),
absolute, 283
due to approximate calculations, 283-288
due to arithmetic operations, 285
computational, 285
irreducible, 283
of the method, 285
relative, 283
Exact methods, 339
Expanding the values of an operator into a series, 190
Extremal(s), 216, 252
Extremum,
necessary condition for, 212
strong, 212
weak, 212

Field(s), 19
central-symmetric, 48
nonstationary, 20
of a point source, 79
scalar, 19
examples of, 19
of sources and sinks, 79
stationary, 20
vector, 19
vector lines of, 79
Flux,
of a vector field, 36-45
of a velocity vector, 39
Formula(s),
complicated quadrature, 326
cubature, 333
Gaussian, 330
Green's, 82, 86
first, 86
second, 86
third, 87
interpolation quadrature, 324

Formula(s),
for numerical differentiation, 318
Ostrogradsky-Gauss, 41-45
in vector form, 48-50
quadrature, 323
rectangular, 324
Simpson's, 324
Stokes', 59
trapezoidal, 324
Function(s),
Green's, of the Dirichlet problem, 97
harmonic, 87-88
properties of, 91
integral representation of, 88
net, 362
of a point source, 98
Functional(s),
dependent on higher-order derivatives, 224-227
extremum of, 208, 211
strong, 212
weak, 212
increment of, 208
linear, 207
maximum of, 211
minimum of, 211
natural boundary conditions for, 247
variation of, 208
Fundamentals of the theory of function approximation, 288-293

Geodesics, 231
Gradient,
in orthogonal curvilinear coordinate system, 23
in a scalar field, 22

Hamiltonian operator, 55
Harmonic, 88
Hausdorff's criterion, 146
Hausdorff's theorem, 147
Heat equation, 108
Hilbert-Schmidt theorem, 191
Hodograph, 9
Hölder's inequality, 116

Inequality(ies),
auxiliary, 114
Cauchy's, 118
Cauchy-Buniakowski, 117
Hölder's, 116
Minkowski's, 118
Schwarz's, 117

Jackson's theorem, 298
Jordan measure, 128

Kernel(s),
Fredholm's iterated, 172

Lagrangian multipliers, 228
Lamé's coefficients, 17
Laplace's equation, 88
vector field, 84
Laplacian (operator), 85
Level lines (surfaces), 20
Limit of a sequence, 122
Line(s),
of force, 25
level, 21
vector, 25
Line integral(s),
of the first kind, 28
of the second kind, 28
Linear manifold, 155
Local bases, 15

Matrix(ces),
banded, 342
characteristic equation of, 346
poorly conditioned, 340
stable, 340
symmetric, 345
unstable, 340
Measure,
Jordan, 128
Lebesgue, 128
Method(s),
Adams', 359
direct, in the calculus of variations, 263
exact, 339
of finding the potential, 72

Method(s),
Galerkin's, 270
Gauss elimination, 340
backward procedure of, 341
forward procedure of, 341
general orthogonalization, 124
iteration, for solving Fredholm's equation, 172
iterative, 339, 344
Kantorovich's, 275
Monte Carlo, 336
net-point, 362
Newton's, for solving systems of equations, 352
Nikolsky, 326
optimization, 330
Ritz', 264-270
Runge-Kutta, 356
Seidel's, 345
of simple iteration, 344
of successive substitution, 344
sweep, 344
Metric spaces, 121-138
complete, 133
isometric, 138
Möbius strip, 13

Nabla, 55
n-dimensional Euclidean space, 124
Neumann's series, 174
Newton's interpolation polynomial, 312
Norm, 153
Numerical
differentiation, 318-322
integration, 322-325

Operator(s),
characteristic value of, 185
completely continuous, 157
continuous, 131, 156
eigenfunction of, 185
eigenvalue of, 185
Fredholm's, 159, 172
monotonically increasing (decreasing), 168
norm of, 180
self-adjoint, 181

Orientation(s),
of a curve, 10
negative, 10
opposite, 10
positive, 10
of a surface, 13

Polynomial(s),
of best approximation, 293
construction of, 299
generalized, 304
Hermite interpolation, 304
interpolation, 304
Lagrange's interpolation, 309
Newton's interpolation, 312
for equal intervals, 313
Potential,
of an electrostatic field, 75
methods of finding, 72
vector, 82-84
Principal error term, 328
Problem(s),
boundary-value, 94
of the choice of a method for an exact or approximate solution, 112
construction, 290
with differential constraints, 233
Dirichlet, 94
of error estimate in an approximate solution, 112
of the existence of the best approximation element, 289
interpolation, 304, 314
isoperimetric, 234-239
of the line of quickest descent, 203
Neumann, 95
solution existence, 112
solution stability, 112
solution uniqueness, 112
Sturm-Liouville, 197
third boundary-value, 262
uniqueness, 289
variation, 203
Procedure,
backward sweep, 344
forward sweep, 344

Reciprocity principle, 240
Rectangular formula, 324
Ritz' method, 264-270
Runge-Kutta method, 356

Schauder's theorem, 163
Sequence(s),
  Cauchy's, 133
  of coordinate functions, 265
  minimizing, 264
  of nested closed balls, 133
Solving the Dirichlet problem with the aid of Green's function, 96
Solving nonlinear equations, 350-354
Space(s),
  Banach, 154
  of continuous functions, 127
  of continuously differentiable functions, 204
  of convergent sequences, 129-130
  Hilbert, 178
    coordinate, 131
    functional, 129
  infinite dimensional, 155
  metric, 121-138
  n-dimensional Euclidean, 124
  n-dimensional vector, 124
  normed linear, 153
  of pth power integrable functions, 128
  separable, 151
  of sequences with convergent series, 130
  of type B, 154
Stokes' theorem, 61
Surface(s),
  orientation of, 13
    negative, 13
    positive, 13
  oriented, 13
  piecewise smooth, 14
Surface integral of the second type, 37
Systems of linear algebraic equations, 339-348

Theorem(s),
  Arzela's, 148
  Banach-Steinhaus, 307
  Bernstein's, 298
  Bolzano-Weierstrass, 145
  Brouwer's, 164
  Chebyshev's, 29
  existence, 299
  Hausdorff, 147
  Hilbert-Schmidt, 191
  Jackson's, 298
  mean-value, 91
  on nested balls, 133-137
  Schauder's, 163
  Stokes', 61
  Vallée-Poussin's, 293

Uniqueness problem, 289
Unit operator, 126

Vandermonde determinant, 309
Variation problem(s),
  connected with Poisson's equation, 260-263
  with fixed boundaries, 215-217
  involving functions of several variables, 250
  with moving boundaries, 241
Vector field(s),
  circulation of, 51
  curl of, 53
  Laplace's, 84-93
  potential, 64-75
  solenoidal, 76-84
Vector function of a scalar argument, 9
Vector lines, 25
Vector tube, 77
  intensity of, 78

Weak and strong extremum of a functional, 211-212
Weight coefficients, 323
Work done by a vector field, 25-30

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book.

We would also be pleased to receive any other suggestions you may wish to make.

Our address is:

Mir Publishers
2 Pervy Rizhsky Pereulok
I-110 GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics

OTHER MIR TITLES

Ya. S. BUGROV, S. M. NIKOLSKY

HIGHER MATHEMATICS

Fundamentals of Linear Algebra and Analytical Geometry

This textbook and two other complementary ones written by the same authors, *Differential and Integral Calculus* and *Differential Equations. Multiple Integrals. Series. Theory of Functions of a Complex Variable*, constitute a course in higher mathematics for engineering students.

Ya. S. BUGROV, S. M. NIKOLSKY

HIGHER MATHEMATICS

Differential and Integral Calculus

The textbook together with two other complementary books written by the same authors, *Fundamentals of Linear Algebra and Analytical Geometry* and *Differential Equations. Multiple Integrals. Series. Theory of Functions of a Complex Variable*, form a course in higher mathematics for engineering students.

Ya. S. BUGROV, S. M. NIKOLSKY

HIGHER MATHEMATICS

Differential Equations. Multiple Integrals. Series. Theory of Functions of a Complex Variable

The present textbook, together with three other books, *Fundamentals of Linear Algebra and Analytical Geometry, Differential and Integral Calculus*, and *A Collection of Problems*, form a four-book series entitled "Higher Mathematics".

The book consists of seven chapters. Chapter 1 is dedicated to ordinary differential equations, and Chapter 2 deals with multiple integrals. Chapter 3 presents a study of vector analysis. Fourier series and Fourier integrals are discussed in Chapter 4. Chapter 5 treats equations of mathematical physics. Chapter 6 contains the theory of functions of a complex variable, and Chapter 7 operational calculus.

Each chapter opens with principal concepts and notions. As a rule, formal proofs of the theorems under consideration are given at the end of the chapter or section. This enables the reader to confine himself, in the case of necessity, to the beginnings of the chapters or sections.

The exercises should be regarded as an integral part of the book. There is a great deal more to be learned from doing the exercises than from a passive reading of the text.

Ya. S. BUGROV, S. M. NIKOLSKY

HIGHER MATHEMATICS

A Collection of Problems

This collection of about 1200 problems has been compiled for the following three textbooks by the same authors: *Fundamentals of Linear Algebra and Analytical Geometry, Differential and Integral Calculus, Differential Equations. Multiple Integrals. Series. Theory of Functions of a Complex Variable*, thus completing a course in higher mathematics for engineering students and forming a four-book series entitled "Higher Mathematics". Academician S. Nikolsky is the author of the two-volume textbook *A Course of Mathematical Analysis* issued in English by Mir Publishers in 1977 and reprinted in 1981.

All the problems are provided with answers, some of the problems are supplied with hints. The book contains many worked problems.

At the beginning of each section references are given indicating the chapters and sections of the above mentioned books where the corresponding theoretical material can be found.

Valentina M Tarpigoreva, Cand.Sc. is an assistant professor in the Department of Higher Mathematics of the Moscow Institute of Electronic Technology. She began teaching in 1951, and in 1967 she defended a thesis entitled "Extremal Problems for Certain Classes of Analytic Functions", for which she received a candidate's degree in the physical and mathematical sciences.

She has over thirty publications, the most significant of which is "Extremal Problems in Hp Classes"

Valentina M. Terpigoreva, Cand.Sc., is an assistant professor in the Department of Higher Mathematics of the Moscow Institute of Electronic Technology. She began teaching in 1951, and in 1967 she defended a thesis entitled "Extremal Problems for Certain Classes of Analytic Functions", for which she received a candidate's degree in the physical and mathematical sciences.

She has over thirty publications, the most significant of which is "Extremal Problems in Hp Classes".

A. EFIMOV, D.Sc. (Phys.-Math.)
Yu. ZOLOTAREV, D.Sc. (Phys.-Math.)
V. TERPIGOREVA, Cand.Sc. (Phys.-Math.)

MATHEMATICAL ANALYSIS
(Advanced Topics)
Part 2

Application of Some Methods of Mathematical and Functional Analysis

Part 2 of this two-part book deals with the fundamentals of vector analysis, the calculus of variations, the elements of functional analysis as applied to the solution of Fredholm's equation, and basic numerical methods. The book was written for engineering students.